Kazimierz Twardowski's conception of imagination. The early-analytical example and contemporary contexts.

Rafał Kur[1]

Abstract. A tribute to the early-analytical provenience of reflections on the phenomenon of the imagination is not only a historical reference. In the absence of a consensus in current theories of imagination, referring to Twardowski can be philosophically refreshing and methodologically inspiring. What's more, it seems that without establishing at least an overall topology of this mental phenomenon, we will not create a formal structure, necessary for logical machine inferences, which would also deal with other issues such as the interpretation of emotions. The problem is not trivial, because the mechanism of imagination is very complex. And that's what Twardowski noticed when proposing a comprehensive (interdisciplinary) approach, so similar at times to some of the current existing proposals.

1 INTRODUCTION

"Images and concepts" by Kazimierz Twardowski [22] appeared in print in 1898 and is one of the first examples of proving mental representations using analytic concepts. Despite passing of time, it is puzzling to see a substantial convergence and validity between current thesis and Twardowski's. Perhaps this is not a coincidence. Twardowski was not only a philosopher, but also a psychologist, with both a descriptive and an empirical approach. His interdisciplinary approach resembled that of a modern cognitive scientist even though he did not have the experimental facilities of modern-day laboratories.

The following article is, on one hand, a tribute to Twardowski and his achievements in analytical philosophy in the context of the contemporary understanding of the imagination, and on the other, will highlight efficiency of methods of proof chosen by appropriate methodology.

What in the context of the contemporary multitude, often mutually exclusive scientific explanations, has an original meaning, and perhaps partly, is it due to an imprecisely defined problem? I am not suggesting an optimistic alternative. I only mark possibilities of approaching this subject while pointing out methodological assumptions in the analytical provenance developed at the Lviv-Warsaw School.

2 TWARDOWSKI'S ACHIEVEMENTS

Twardowski's he concept of imagination is above all astonishing by its timeliness which has motivated its recall and reinterpretation in this article. Even more so, as is the case of the contemporary lack of consensus on the interpretation of the phenomenon of imagination which additionally makes the research quite refreshing to present and because the theories of the imagination are treated as exemplifications of mental representations. Moreover, research disputes around the issue of imagination reveal a broader dispute about the mechanisms of formation of representation. This is important, because in the current various existing interpretations analytic typologies can help in ordering the methodological levels of the conglomerate of meanings of representation. Contemporary discussions offer detailed mechanistic solutions, which has additionally differentiated the explanations of the issue and even invited arguments for the non existence of representation (anti-representationalism). However, the complete exclusion of representation from cognition in humans (and some animals), would imply the lack of connection of the mental with the external world. Morover, recently the term of representation is widely using in many theories, from humanities, until exact sciences. Therefore, this work supports the existence and the possibility of creation of various forms of representation.

The novel way in which Twardowski's teacher presented the nature of the mental representation of objects sparked his interest. Brentano moved away from Cartesian dualism in which the mind (perceived as thinking or consciousness) was a model reflecting reality. Brentano broke down the Kant's knowing process into representations and judgements. He introduced intentional acts which produce content, while content depicts objects outside of consciousness. Thus, he developed the notion the linkage of a subject with the world, and at the same time initiated a new field for speculation over consciousness. Consequently, Twardowski's original contribution was to distinguish the subject and its representation, thanks to the psychological development of the theory of intentionality and emphasis on the causative activity of the subject (activities/outcomes). These types of arguments were also further developed by his students at the Lviv-Warsaw School. On the other hand, phenomenology (Husserl and students) intensively addressed the philosophical approach to consciousness and intentionality. Mental representations (before known as mental imagery) have gained a new, deeper meaning, whose status by examining various exemplifications, is still a subject of dispute. Twardowski initiated this approach in his work, Zur Lehre vom Inhalt und Gegenstand der Vorstellungen. Eine psychologische Untersuchung, "On the Content and Object of Presentations. A psychological examination" (1894, [24]), where he differentiated Brentan's transcendent from the immanent object of intention. The former denotes an object that is independent of our consciousness, for example a visible object. The latter views the object as a conscious product of this act, more precisely its content. This distinction is important, because

Dept. of Philosophy, Jagiellonian Univ., Grodzka 52, 31-044 Krakow, Poland. Email: rafalkur@yahoo.pl

content deepens the meaning of representation, for example when imagining an object no longer being perceived or an abstract form. Therefore, the distinction between "imagination" and "concept" was, a natural consequence of a refined understanding of content produced during visual (perceptions, representations) and non-visual (concepts, judgments) presentations.

3 THE IMAGERY DEBATE

Twardowski's voice in the discussion on the nature of concepts resulted primarily from attempts to understand objects which we cannot imagine, and which can only be replaced by concepts (infinity, quant, round square, God, etc.). For this reason, Twardowski needed to organise his theories and make a critical selection, then create a general and compact theory analysing types of concepts. A general enough theory to cover all cases of conceiving an object with the help of concepts. Accordingly, he tried to determine limits of "the power of the imagination" beyond which concepts including abstract (rational) objects can exist. The analysis of the established boundaries led Twardowski to an interdisciplinary theory of representation emphasizing the interdependence of imagination and concept, a synthesis. However, "the concept is a representation of an object that consists of a similar imagined object and one or several imagined judgments relating to the imagined object" [22].

The research topics from the Lviv-Warsaw School resemble contemporary interdisciplinary research on mental phenomena. The certainty of some diagnoses of researchers from the School results mainly from the selection of specific methods of exact and natural sciences. It was also a good base for the then emerging psychology that inherited the philosophical issues/problems of the mind, developing and applying its methods, both theoretical (descriptive) and experimental (at the level of physiology). Twardowski's attempt to understand mental representations interestingly turned into a contemporary psychophysical dispute (the body-mind problem). Especially in cognitive psychology, this subject gained special significance due to empirical verifications. Described as the *imagery debate*, it concerns, in fact, the problem of representation and more precisely the mechanisms of coding information in the human cognitive system. As a result of this dispute, since the 1970s it has been cited by both philosophers of the mind and cognitivists. The best-known proposals oscillate around two major competing concepts. Admittedly, the two major competing concepts contain numerous complementary elements, accrued over decades of discussion and supported by advanced experiments. On the one hand, we have the image concept (i.e. analog, visual) postulating that mental images resemble images of real objects, perceived objects and concepts referring to them are represented in the mind in the same form (i.e. specific size and spatial position). These properties are captured directly in the image and not represented in a symbolic (semantic) manner. Representatives: Kosslyn [10, 11, 12]; Shepard, Metzler [21]; Francuz [3]. On the other hand, the proposition which indicates that mental representations are a collection of judgments (i.e. propositional) about the relations between symbols, encoded in the memory as tacit knowledge. Representatives: Pylyshyn [17, 18, 19, 20], Anderson and Bower [2].

The main problem of the image approach lies in the unsatisfactory explanation of the abstract concept representation. Twardowski noticed this a hundred years earlier, but without sufficient tools, he consciously abandoned further analyses of this problem. He proposed, however, and presented possibilities of resolving this problem through sets of claims, beliefs, or judgments. Thus, his approach resembles that of Pylyshyn, the main opponent of the visual nature of representation. Twardowski draws attention, for instance, to the possibility of a double grammatical construction in which the word "think" occurs. One could think about a certain event (imagining - using images and concepts) and think that an event was inevitable (expression of beliefs). Twardowski did not think that the representation itself is but a judgment. In order for the judgment form, it must include a mental act of recognition or rejection, confirmation/sanction or denial, also containing an emotional (internal sensations, physiological and behavioural component) correlations [4, 14, 15].

4 COMMON PARALLELS

It seems that imagery debate is nothing more than an expansion of Twardowski dilemma through experiments. The indirect placement of the Polish philosopher in this discussion results from the surprising accuracy of his analytic deduction, that led him to rationally suspend certain issues and assume an intermediate position. Due to the role of Twardowski's judgments, one could straightaway assume similarity with Pylyshyn's approach, but it should be noted that Twardowski attributed judgments a complementary role, rather than a primary one. Reinterpreting Twardowski in the context of contemporary theories of the imagination is also an indication of qualities of the epistemological tradition from which contemporary reductionist theories seem to be moving away. For instance, the unilateral view that perception is a cognitive act in relation to physical objects narrows the understanding of perception. As perception's building blocks (structure) consist not only of external sensations, but also of internal, resulting from e.g. fun, pain, sadness, love, etc. Twardowski, as the successor of Brentanism, and a witness of expanding behaviorism, did not accept only using "hard methods" in the complex system that is cognition, which currently reflects eliminationism (Particia and Paul Churchland).

Mechanistic (computational) interpretations of the representation are also not completely satisfactory. An interesting example is a recently created model of the Neuronal Turing Machine (NTM), which made us realize that the neurodynamics of the brain cannot be replicated merely by operationalizing data, because the human mind is more than a just a "Turing Machine".

In the absence of unanimity (unifying theory), nothing is more necessary than a sensible methodology and a moderate approach. Perhaps that is why, in cognitive psychology, Alan Paivio's dual coding theory is quite often invoked (cited). Though, in the wider cognitive view, it seems that the closest to Twardowski were the compositional and naturalistic concepts of the mind of Jerry Fodor.

Allan Paivio [16, 17] assumes the existence of two separate information processing systems. The non-verbal system responsible for coding information in the form of multimodal

patterns of activation of the network of neurons associated with perception, which are then used to simulate the perception of objects. And the language system responsible for coding information in the form of relations between symbols of different strengths of association organized hierarchically as nodes in the semantic network, which can connect with each other and with objects represented in the non-verbal system. From the Twardowski point of view, an interesting fact is that knowledge coded in the language system is contained in the network of relations between symbols and not in the symbols themselves. A single node of such a network can often be identified with a single category, associated on one side with the corresponding word, and on the other with a certain class of objects encoded in a non-verbal (image) system.

Behavioral experiments conducted by Paivio indicate that Twardowski assumption were correct. Due to some conclusions Canadian psychologist, for example in the case of too much of a categorical separation of coding content of specific concepts in both systems, the content of abstract concepts only in the language system, to counteract this, Twardowski's thesis, for example, on the fluidity between abstract and specific concepts becomes very useful. The dual coding theory does not contradict the results of experiments that support opposition positions, it also seems to be in line with the current state of neurobiological knowledge.

Why did Twardowski consider images and concepts to be the most important form of representation and treat them as complementary? In the light of today's research, can you keep his proposal? Probably yes. The achievements of cognitive science bring us closer to different proposals. Contemporary experiments in cognitive psychology do not definitively admit any of the parties to the dispute, although it is currently noticeable that the "visual" approach is more popular. Using Twardowski's work as inspiration, I would argue that the complex content (neuro-dynamic format of information) of presentations requires an interdisciplinary approach. Moreover, the mereological nature of Twardowski's assertions is also a methodological (formal) clue to the dynamic and diverse representational resources [1].

On the other hand, the inclusion of other correlations (e.g. emotional) emphasizes the proto-cognitive character of Twardowski's considerations. Twardowski wrote: "If the idea is not a renewed insight in general, nor a simple recreation of sensations, there is nothing else but to seek them in the very synthesis of sensations (...). As a synthesis of sensations, an idea is based on sensations - whether immediate or refreshed - though it is not a simple recreation of them; it can therefore be based on any impressions, as long as a proper whole can be made of them. (...) Imagination, therefore, is to a sensation, like the whole to a part. One could ask what kind of synthesis is the one in which the sensations are arranged to create the images/imagination. But psychology has not been able to and probably never will be able to formulate an answer to this question" [22].

The characteristic arguments concerning the synthesis and interdependence of elements of the representations presented by Twardowski resemble proposals of Jerry Fodor [5, 6, 7, 8, 9], who postulates that the types of mental representations are of a compositional nature, dividing them into two basic types, i.e. linguistic (conceptualized) and iconic (conceptualized). In the case of iconic or "visual", these need empirical evidence. Analyzing the differences and similarities between linguistic and iconic representations, Fodor arrives to similar conclusions as Twardowski's did a century ago. For example, the lack of a logical form of iconic representation, a characteristic relationship of parts to the whole that complement each other at a general level. Fodor's naturalistic idea derived from the criticism of the inferential position of Frege, who unnecessarily - according to Fodor - associated methods of presenting objects only with the meaning of language expressions. Fodor from the 1960s, while working with Noam Chomski, he began opposing behaviorism, when it turned out that internal representations could explain many more properties of cognition - from the laws of perception to the cognitive foundations of logic and language. The content of the representation is a complex creation connected with the cognitive system with many causal links, including semantic properties. Associations of this type according to Fodor explain the productivity and regularity of our thoughts. The reference-based semantics makes it possible to refine (individualize) concepts. This type of inference is similar to Twardowski's method, whose conceptual apparatus seems suitable - after making some modifications - to the role of a peacemaker between reductionistic neuroscience and speculative philosophy of the mind. Additionally, Fodor argues that the computational nature of mental processes brings the philosophy of mind closer to cognitive science based on IT methods. Interestingly, Fodor has no commentary on the analysis of the factors defining concepts. He admits to lack some element in the theory representing the mind and suggests only a conceptual framework for a future theory.

5 CONCLUSION

It was obvious to Twardowski that the mind, the subject of the study of philosophers and psychologists, is associated with the biological brain. The problem he could not solve, and which, in fact, exists to this day, was the lack of obvious details of the relationship. This psychophysical dilemma Twardowski tried to explain, by introducing the concept of a function. "The mental activity is reliably a function of the brain in the first sense of the word, because certain changes taking place in the brain involve changes in mental activity. One can not call the atoll of the mental activity the function of the brain in the second of the meanings quoted. There is no evidence to suggest that mental activity is carried out completely and exclusively by the brain" [25]. Mental activities are not isolated from the brain, nor are they detached from external reality. Twardowski, through the analysis of various activities emphasizes how entangled we are with the world, and thus cognition is embodied. However, he could not study the source of the psychophysical. Nowadays, even neuroscientists are reserved in explaining these issues; numerous experiments usually further complicate the things. This is why cognitive science is developing so dynamically. On the one hand, it makes use of evidence from cognitive psychology, and on the other, extensively uses the theoretical assumptions of analytical philosophy of the mind.

Understanding the relationships between the neurobiological (physical) processes of the brain and mental reactions is still the *body-mind problem*. Among various approaches (reductionism, epiphenomenalism, dualism, etc.) Twardowski's proposals in light

of contemporary research are an opportunity to recall some concepts of the Lviv-Warsaw School, including the moderate and interdisciplinary (also known as comprehensive, mixed, cross-domain) methodological proposals in the study of mental representations, as a reaction to the overly reductionist trends in cognitive science.

REFERENCES

[1] L. Albertazzi. The Primitives of Presentation. Wholes, Parts and Psychophysics. In: The Dawn of Cognitive Science. Early European Contributors. L. Albertazzi (Ed.). Synthese Library 295. Springer, Netherlands (2001).

[2] J.R. Anderson, G. Bower. Human associative memory. Wiley, New York, USA (1973).

[3] R. Tadeusiewicz. Awangarda sztucznej inteligencji - maszyny które potrafią same tworzyć nowe pojęcia". In: Pojęcia. Jak reprezentujemy i kategoryzujemy świat, [The avant-garde of artificial intelligence - machines that can create new concepts themselves. In: Concepts. How we represent and categorize the world]. J. Bremer, A. Chuderski (Eds.). Universitas, Krakow, Poland (2011).

[4] A. Damasio. Descartes' Error: Emotion, Reason, and the Human Brain. Putnam Publishing (1994).

[5] J.A. Fodor. Representations. The MIT Press, Cambridge (1981).

[6] J.A. Fodor. The modularity of mind. MIT Press, Cambridge (1983).

[7]J.A.Fodor, Z.W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28, 3-71 (1988).

[8] J.A. Fodor. Concepts: Where Cognitive Science Went Wrong. Oxford Cognitive Science Series (1998).

[9] J.A. Fodor. LOT 2: The language of Thought Revisited. Oxford University Press, USA (2008).

[10] S.M. Kosslyn. The medium and the message in mental imagery: A theory. *Psychological Review*, 88: 46-65 (1981).

[11] S.M. Kosslyn. Image and brain: The resolution of the imagery debate. MIT Press, Cambridge (1994).

[12] S.M. Kosslyn., and S.P. Shwartz. A simulation of visual imagery. *Cognitive Science*, 1: 265-295 (1977).

[13] S.M. Kosslyn, W.I. Thompson, G. Ganis. The case for mental imagery, Oxford University Press (2006).

[14] J.E. LeDoux. Emotion circuits in the brain. *Annual Review of Neuroscience*, 23: 155-184 (2000).

[15] T.I. Lubbard, I. Getz. Emotions, metaphor and the creative process. *Creativity Research Journal*, 10: 285-302 (1997).

[16] A. Paivio. Mental representations. A dual coding approach. Oxford University Press, New York (1990).

[17] A. Paivio. Mind and Its Evolution: A Dual Coding Theoretical Approach. Taylor & Francis Group, New York (2006).

[18] Z.W. Pylyshyn. What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Biuletin*, 80: 1-24 (1973).

[18] Z.W. Pylyshyn. The imagery debate: analogue versus tacit knowledge. *Psychological Review*, 88 (1981).

[19] Z.W. Pylyshyn. Seeing and Visualizing: It's Not What You Think, MIT Press (2004).

[20] Z.W. Pylyshyn. Things and Places: How the Mind Connects with the World, MIT Press (2007).

[21] R.N. Shepard, J. Metzlez. Mental rotation of three-dimensional objects. *Science*, 171: 701-703 (1971).

[22] K. Twardowski, Wyobrażenia i pojęcia [Images and concepts]. Lwów [Lviv] (1898). In: Wybrane pisma filozoficzne [Selected philosophical papers]. PWN, Warszawa [Warsaw]. pp. 114-197. (1965).

[23] K. Twardowski. O czynnościach i wytworach. Kilka uwag z pogranicza psychologii, gramatyki i logiki [On Actions and Products. Some remarks from borderline of psychology, grammar and logic] Lwów [Lviv] (1927). In: Wybrane pisma filozoficzne [Selected philosophical papers]. PWN, Warszawa [Warsaw]. pp. 217-240 (1965).

[24] K. Twardowski. O treści i przedmiocie przedstawień [On the content and object of presentations]. In: Wybrane pisma filozoficzne [Selected philosophical papers]. PWN, Warszawa [Warsaw]. pp. 3-91 (1965). Transl. by I. Dąmbska from original: Zur Lehre vom Inhalt und Gegenstand der Vorstellungen. Eine psychologische Untersuchung. Wien (1894).

[25] K. Twardowski. Psychologia wobec filozofii i fizjologii. Lwów [Lviv] (1927). In: Wybrane pisma filozoficzne [Selected philosophical papers]. PWN, Warszawa [Warsaw]. pp. 92-113 (1965).

[26] J. Woleński. Logic and Philosophy in the Lvov-Warsaw School. Dordrecht, Kluwer (1989).

[27] S. Lapointe, J. Woleński, M. Mathieu, W. Miśkiewicz. The Golden Age of Polish Philosophy. Kazimierz Twardowski's Philosophical Legacy. Springer, Dordrecht (2009).

Glanville's 'Black Box': what can an Observer know?

Lance Nizami[1]

Abstract. This paper concerns the 'Black Box'. It is not the engineer's 'black box' that can be opened to reveal its mechanism, but rather, one whose operations are inferred through input from (and output to) a companion 'observer'. We are observers ourselves, and we attempt to understand *minds* through interaction with their host organisms. To this end, Ranulph Glanville, Professor of Design, followed the cyberneticist W. Ross Ashby in elaborating the Black Box. The Black Box and its observer together form a system having different properties than either component alone, making it a 'greater' Black Box to any further-external observer. However, Glanville offers conflicting accounts of how 'far' into this 'greater' box a further-external observer can probe. At first (1982), the further-external observer interacts directly with the core Black Box while ignoring that Box's immediate observer. But in later accounts, the greater Black Box is unitary. Glanville does not explain this discrepancy. Nonetheless, a firm resolution is crucial to understanding 'Black Boxes', so one is offered here. It uses von Foerster's 'machines', abstract entities having mechanoelectrical bases, just like putative Black Boxes. Von Foerster follows Turing, E.F. Moore, and Ashby in recognizing archetype machines that he calls 'Trivial' (predictable) and 'Non-Trivial' (non-predictable). Indeed, early-on Glanville treats the core Black Box and its observer as Trivial Machines, that gradually 'whiten' (illuminate) each other though input and output, becoming 'white boxes'. Later, however, Glanville treats them as Non-Trivial Machines, that never fully 'whiten'. Non-Trivial Machines are the only true Black Boxes. But Non-Trivial Machines can be concatenated from Trivial Machines. Hence, the utter core of any 'greater' Black Box (a Non-Trivial Machine) may involve two (or more) White Boxes (Trivial Machines). White Boxes may be the ultimate source of the mind.

[1] Independent Research Scholar, Palo Alto, CA 94306, USA. Email: nizamii2@att.net

1 INTRODUCTION

One dream of Artificial Intelligence is to exactly mimic natural intelligence. Natural intelligence is examined by observing the behaviors that imply a *mind* (unlike mere physiological reflexes). Minds presumably exist within animals having recognizable brains. (Whether other species have minds will not be debated.)

Of course, behavior can be difficult to quantify, especially when experimental-research subjects cannot 'report'. An example of reporting is the confirming of particular sensations evoked by stimuli [1-3]. In animals, primitive reporting (Yes/No, Left/Right, etc.) can be painstakingly conditioned. But conditioning may not be feasible (or ethical) for human infants. Nonetheless, we desire to establish infants' sensory abilities, in order to detect defects [1-3]. But, by-and-large, the animal or the infant is a 'Black Box' to the observer of behavior [1-3]. The reason for the capital B's will soon be explained.

Sensations are a feature of the *mind*. Here, the attempts to understand a mind through interaction with its host organism are placed in relation to the notions of the Black Box and its 'observer' as proselytized by Ranulph Glanville [4-8]. Glanville was a Professor of Design and a champion of Second-Order Cybernetics. Glanville notes [6] that much of his discourse on the Black Box originates in the writings of W. Ross Ashby. Hence, we begin with Ashby. Ashby devotes a chapter to the Black Box in his book *An Introduction to Cybernetics* (1956), a book cited over 11,000 times (GoogleScholar). (For a brief summary of Ashby's importance to science, see [9].) Ashby's 1961 edition is more readily available, and is cited here [10].

2 THE BLACK BOX AND ITS OBSERVER

The 'black box' of an engineer or a physicist is a physical object that can be opened, letting its operation be comprehended. If un-openable, however, this 'machine' becomes a Black Box [10], understood only through inputs given by, and outputs noted by, an observer [10]. Indeed, the input/output cycle may never reveal the Black Box's mechanism; a *mechanical* basis, for example, may be indistinguishable from an *electrical* one [10]. Ashby [10] gives examples, noting that we can "Cover the central parts of the mechanism and the two machines are indistinguishable throughout an infinite number of tests applied. Machines can thus show the profoundest similarities in behavior while being, from other points of view, utterly dissimilar" ([10], p. 96).

Following Ashby, we might imagine machines that consist of both mechanical *and* electrical components, *mechanoelectrical* 'systems' whose actual mechanisms are indistinguishable, one from another, through input and output. As such, the mechanisms become irrelevant. Glanville takes this logic to its limit: "You cannot see inside the Black Box (there is nothing to see: there is nothing there—it is an *explanatory principle*)" ([5], p. 2; italics added). That is, "Our Black Box is not a physical object, but a concept … It has no substance, and so can neither be opened, nor does it have an inside" ([8], p. 154). Even so, Glanville states that it has a *mechanism* [4, 6-8].

Glanville's Black Box may sound suspiciously like a *mind*. After all, no-one can directly observe their own mind, or anybody/anything else's; *"mind" is an explanatory principle* for what we call 'behavior'. Glanville's work [4-8] therefore deserves further scrutiny. Unfortunately, his principal exposition [4] requires clarification, as will be explained. Glanville later attempts clarification [5-8], but falls short. The present paper provides the missing details. Provocative insights emerge.

3 'WHITENING' THE BLACK BOX

Let us clarify Glanville's notion of the Black Box as a "phenomenon" or "principle" or "concept". First, let us assume that the Black Box is spatially located. This forces another assumption, namely, that wherever the location, there must be a mechanoelectrical system that is the basis for – that *produces* – the Black Box. For example, the brain with its extended network of neurons and blood vessels indisputably *produces* the mind, whose existence is evident through non-reflexive *behavior*.

('Reflexive' behavior would include, for example, the jerk of the lower leg when the knee is tapped by a physician, or the tendency of some single-celled organisms to move towards light.). *The mind is not independent of its host body; likewise, the Black Box is not independent of its mechanoelectrical basis*.

Figure 1 schematizes the Black Box and its observer. The observer makes inferences about the Black Box by presenting stimuli, the inputs, and recording the Box's consequent responses, the outputs [4-8, 10]. According to Glanville ([4], p. 1), the observer thereby obtains a "functional description" of the Black Box: "The 'functional description' … describes how the observer understands the action of the Black Box" ([5], p. 2). That is, the Black Box is 'whitened' [4]. Practical examples of 'whitening' through input/output might include an Experimental Psychologist studying the behavior of a human or an animal, or a Physiologist making a noninvasive electrical recording [1, 11].

4 OBSERVER AS BLACK BOX, BLACK BOX AS OBSERVER

'Whitening' of the Black Box becomes more intriguing yet. Glanville [4, 5, 7, 8] declares that the *observer* can be considered a Black Box, from the *Black-Box*'s viewpoint. Consider that an *output* of the Black Box is an *input to* its observer; likewise, an input to the Black Box is an *output from* the observer. Hence, "we come to assume that the Black Box also makes a functional description of its interaction with the observer" ([5], p. 2). The Black Box 'whitens' its observer, by *acting as* an observer [4].

Consider the following examples. Imagine the mind as a Black Box, probed through input and output. We call this action Psychiatry or Psychology. But each Psychiatrist or Psychologist has their own mind, a Black Box. Those particular Black Boxes regulate everything that those observers say and do; the observers are therefore now Black Boxes. And indeed, Moore [12] and Ashby [10] both imply that a *Psychiatrist* and a patient are interacting Black Boxes. When the Psychiatrist (or the Psychologist) probes the patient (or the research subject), e*ach participant* (if awake and aware) *is an observer, who regards the other as a Black Box*. Such interaction implies a *system*.

5 BLACK BOX + OBSERVER = 'SYSTEM': INSIDE EVERY WHITE BOX THERE ARE TWO BLACK BOXES TRYING TO GET OUT

Ashby analyzes experiments as follows: "By thus acting on the Box, and by allowing the Box to affect him and his recording apparatus, the experimenter is coupling himself to the Box, so that the two together form a *system* with feedback" ([10], p. 87; italics added). That is, experimenter and Box each "feed back" to the other, each becoming both observer and Black Box. A BlackBox/observer *system* has different properties than either the Black Box or the observer alone, or so Glanville implies: "The Black Box and the observer act together to constitute a (new) whole" ([7], p. 1; see [8], p. 161). This he calls the *white box* [4].

Figure 2 schematizes the 'white box'. If now the observer himself is taken to be a Black Box, then the title of Glanville's paper of 1982 [4] becomes comprehensible: "Inside every White Box there are two Black Boxes trying to get out". According to Glanville [4, 7, 8] the White Box, as a system, is nonetheless 'black' to any *further-external* observer.

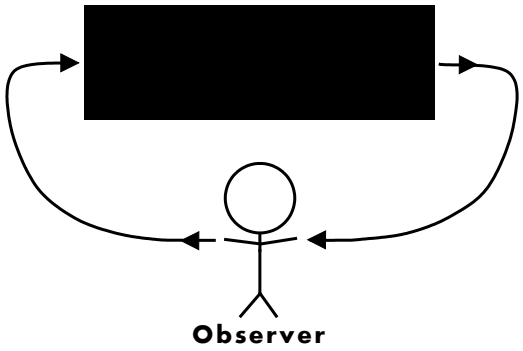

Figure 1. (After [4]) The Glanville notion of the Black Box and its observer. The observer sends inputs to the Box, and receives outputs from it.

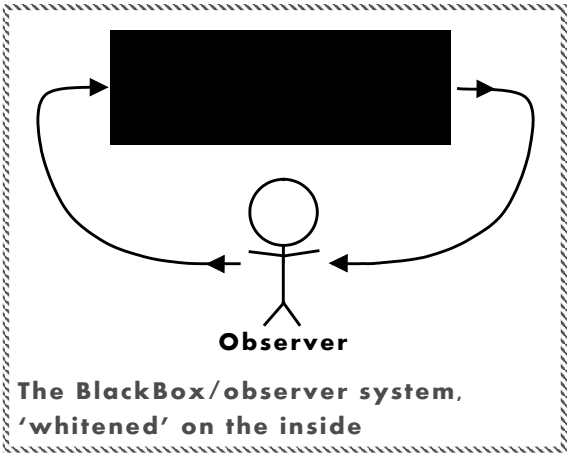

Figure 2. (After [4]) The Black Box and its observer mutually 'whiten' through interaction, making a 'system' (dashed boundary) that is 'whitened' inside.

6 HOW WOULD A FURTHER-EXTERNAL OBSERVER INTERACT WITH THE SYSTEM?

We have assumed that the Black Box is the *product* of a mechanism. The pictured boundary of any Black Box (and any consequent White Box) is *operational*, not physical. Where, therefore, can inputs from a further-external observer go within the BlackBox/observer system? Do they go directly to the core Black Box? Or to the [original] observer? Or, somehow, to both? The answer presumably tells us where consequent outputs originate from too. Figures 3 and 4 illustrate two possibilities.

Figure 3 illustrates the further-external observer's inputs as going *straight through* the boundary of the BlackBox/observer system, right up to the edge of the core Black Box itself, without interacting with the core Black Box's observer. The latter persona is ignored, as if the further-external observer recognizes his presence and behavior. Now consider the contrary situation. Figure 4 (after [8], p. 164) shows the core BlackBox/observer system *not* being penetrated by the input-and-output pathways to

the further-external observer. Indeed, Glanville [7] implies that in "a recursion of Black Boxes (and observers)", *none of the observers know of each other's existence*. But Glanville [8] later fails to be definitive about this. Indeed, he provides no rationale for the discrepancy between his approach of 1982 [4] and his approach of 2009 [7, 8]. And he can no longer provide one [13]. Consequently, the present author attempts the task. Important insights emerge, but first, some background is needed.

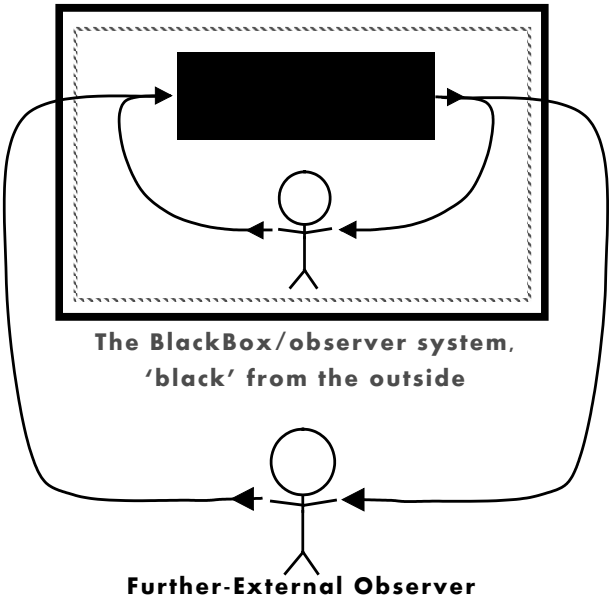

Figure 3. (After [4]) The core BlackBox/observer system as a Black Box which is penetrable by the input/output paths from/to a further-external observer.

7 INTERIM (1): 'TRIVIAL' MACHINES

Consider the *machine*, introduced in Section 2 as a mechanoelectrical device. Henceforth, 'machine' will be used in a different way. As von Foerster ([14], p. 207) states, "The term 'machine' in this context refers to *well-defined functional properties of an abstract entity* rather than to an assembly of cogwheels, buttons and levers, although such assemblies may represent embodiments of these abstract functional entities [i.e., the 'machines']" (italics added). By this definition, an abstract entity that is a *product* of a mechanoelectrical basis – i.e., an abstract entity such as the Glanville Black Box – is a 'machine'.

Von Foerster recognizes two types of machines: Trivial, and Non-Trivial. He explains: "A *trivial* machine is characterized by a one-to-one relationship between its 'input' (stimulus, cause) and its 'output' (response, effect). This invariant *relationship* is 'the machine'. Since this relationship is determined once and for all, this is a deterministic system; and since an output once observed for a given input will be the same for the same input given later, this is also a *predictable* system" ([14], p. 208; italics added). Algebra-wise, von Foerster explains that for input **x** and output **y**, "a **y** once observed for a given **x** will be the same for the same **x** given later" ([15], p. 9). That is, "one simply has to record for each given **x** the corresponding **y**. This record is then 'the machine'" ([15], p. 10).

Von Foerster ([15], p. 10) provides an example of a Trivial Machine, in the form of a table which assigns an output **y** to each of four inputs **x**. The **x**'s are the letters A, U, S, and T, and the respective outputs **y** are 0, 1, 1, and 0. Von Foerster ([14], p. 208) notes that "All machines [that] we construct and buy are, hopefully, trivial machines", that is, *predictable* ones.

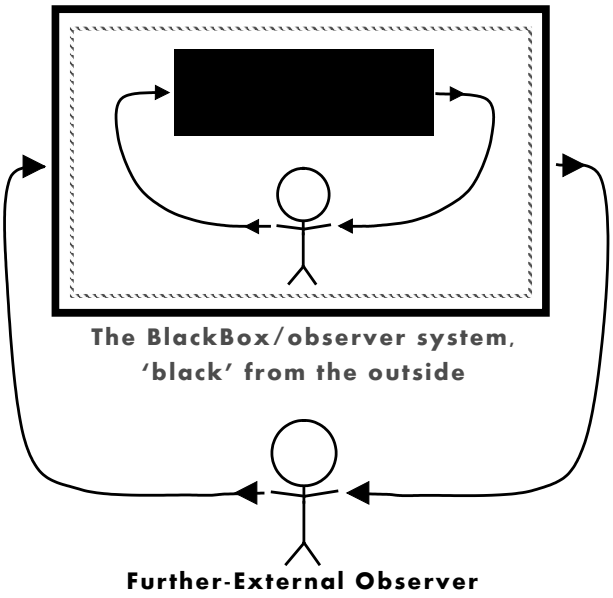

Figure 4. (After [8]) Compare to Fig. 3. The core BlackBox/observer system as a Black Box which is *not* penetrable by the input/output paths from/to a further-external observer.

8 INTERIM (2): INTERNAL 'STATES'

Of course, in reality, *not all machines are Trivial*. A *physical* machine, as Ashby [10] points out, can have internal conditions or configurations, which Ashby calls 'states'. We will assume that so, too, can the *products* of physical machines, namely, *conceptual* machines such as Black Boxes. Ashby notes "that certain states of the Box cannot be returned to at will", which he declares "is very common in practice. Such states will be called **inaccessible**" (all from [10], p. 92; original boldface). Ashby continues: "Essentially the same phenomenon occurs when experiments are conducted on an organism that *learns*; for as time goes on it leaves its 'unsophisticated' initial state, and no simple manipulation can get it back to this state" ([10], p. 92; italics added). *Learning* presumably means changes in abilities and knowledge, that are reflected in changes of behavior.

Here, *mind* is *machine* is *Black Box*. Learning is presumably concurrent with changes in the mind's mechanoelectrical basis, the brain; changes in brain-states manifest as changes in mind-states. Thanks to learning, our response to a stimulus can differ from our previous response, and in unexpected ways. For a stimulus that is a *question*, for example, von Foerster ([14], p. 311) notes that a child can offer a correct [carefully trained]

answer, or a correct but unexpected answer, or an answer that is intentionally capricious. *Minds are not Trivial Machines*.

We must therefore ask whether the observer of any Black Box can give input, and record output, without changing the Box's possible output to the next input. That is, can the Black Box be *observed* without being *perturbed*? Likewise, can a Black Box's output be observed by, but not perturb, the observer?

9 INTERIM (3): SEQUENTIAL MACHINES

There are conceivably perturbable 'machines'. Such devices were envisioned long before Ashby [10]. Indeed, Turing [16] conceives of a machine whose input and/or output can change the response to the next input. Turing describes the machine only in terms of its process. The machine has a finite number of internal states, called 'conditions' or 'configurations'. The machine accepts an input in the form of a continuous tape, divided into equal segments, each containing a symbol or being blank. The machine scans one tape segment at a time; the scanned symbol (or the blank), along with the machine's current configuration, altogether determine the impending response. That response can include erasing a symbol from the tape; or printing (or not), on a blank segment of the tape, a symbol consisting of a digit (0 or 1) or some other symbol; or shifting the tape one segment to the left or one segment to the right [16].

Turing's machine exemplified what came to be known as *sequential machines*. One class of them was described by E.F. Moore [12]. He, like Turing, used operational descriptions: "The state that the machine will be in at a given time [the 'current state'] depends only on its state at the previous time and the previous input symbol. The output symbol at a given time depends only on the current state of the machine" ([12], p. 133). That is, an input evokes an output, which is nonetheless determined *only* by the present internal state. That state then changes to another state, that is determined by the *input*.

Moore [12] provides an example of a sequential machine, in the form of two tables that relate the inputs, outputs, and internal states ([12], p. 134). Let us call the inputs **x**. Moore's inputs are also the possible outputs, but for the sake of distinction, let us call the outputs **y**. One of Moore's tables shows the "present output" **y** of the machine, as a function of the "present state", call it **z**. Let this relation be called **y**=**F**(**z**) and be satisfied by an 'Output Generator'. Moore's second table shows "the present state of the machine … as a function of the previous state and the previous input" ([12] p. 134). Let us call Moore's "previous state" $\mathbf{z}_{-1}$ and the "previous input" $\mathbf{x}_{-1}$. Let us use $\mathbf{z}'$ for the state of **z** which occurs *after* **y** is output. Let $\mathbf{z}_{-1}$, **z**, and $\mathbf{z}'$ be determined by the 'State Generator' **Z**, and express **Z** in terms of **z** rather than $\mathbf{z}_{-1}$. Moore [12] uses four possible internal states, called $\mathbf{q}_1$, $\mathbf{q}_2$, $\mathbf{q}_3$, and $\mathbf{q}_4$, and two possible inputs, **x** = 0 or **x** = 1. All of this notation may seem awkward, but it is consistent with the work of von Foerster [14, 15], continued below.

Table 1 shows a re-arrangement of Moore's two tables into five smaller tables, four of which show **z** as a function of $\mathbf{x}_{-1}$ for the four possible values of $\mathbf{z}_{-1}$($\mathbf{q}_1$, $\mathbf{q}_2$, $\mathbf{q}_3$, and $\mathbf{q}_4$), the remaining table showing **y** as a function of **z**.

As an example of how the Moore sequential machine works, note that $\mathbf{z} = \mathbf{q}_4$ could have arisen from $\mathbf{z}_{-1} = \mathbf{q}_3$ and $\mathbf{x}_{-1}$ = 0 or 1, or from $\mathbf{z}_{-1} = \mathbf{q}_1$ and $\mathbf{x}_{-1}$ = 0. Regardless, an input **x** = 0 *or* **x** = 1 now evokes the output **y** = 1, after which $\mathbf{x}_{-1}$ = 0 or $\mathbf{x}_{-1}$ = 1, respectively, and $\mathbf{z}_{-1} = \mathbf{q}_4$, leading to a new internal state $\mathbf{z}' = \mathbf{q}_2$. A subsequent input **x** = 0 *or* **x** = 1 will result in **y** = 0, and so on.

$\mathbf{z}_{-1} = \mathbf{q}_1$	
$\mathbf{x}_{-1}$	**z**
0	q_4
1	q_3

$\mathbf{z}_{-1} = \mathbf{q}_2$	
$\mathbf{x}_{-1}$	**z**
0	q_1
1	q_3

$\mathbf{z}_{-1} = \mathbf{q}_3$	
$\mathbf{x}_{-1}$	**z**
0	q_4
1	q_4

$\mathbf{z}_{-1} = \mathbf{q}_4$	
$\mathbf{x}_{-1}$	**z**
0	q_2
1	q_2

z	**y**
q_1	0
q_2	0
q_3	0
q_4	1

Table 1. Relations in Moore's example of a sequential machine [12]. The rightmost table describes the Output Generator; the other four tables describe the State Generator, for internal states $\mathbf{q}_1$, $\mathbf{q}_2$, $\mathbf{q}_3$, or $\mathbf{q}_4$.

Note well that, in Moore's scheme, a particular output can result from different internal states; and a particular internal state can result from different *inputs*. Note equally well that Moore's two tables are *unchanging*. That is, what we presently call the State Generator and the Output Generator are deterministic (i.e., non-random) *and* they are predictable, insofar as an outside observer supplying input and recording output can gain increasing confidence about each Generator's operating rules. *Both Generators are Trivial Machines.*

But the concatenation of two Trivial Machines can be non-trivial, i.e., non-predictable; the whole is more than the sum of the parts. This wholism is called 'emergence' [1, 17]. How would an *observer* of the sequential machine (not its *maker*) gain the data to fill Moore's two tables? Moore introduces "a somewhat artificial restriction that will be imposed on the action of the experimenter. He is not allowed to open up the machine and look at the parts to see what they are and how they are interconnected" ([12], p. 132). That is, "the machines under consideration are always just what are sometimes called 'black boxes', described in terms of their inputs and outputs, but no internal construction information can be gained" ([12], p. 132).

Moore himself offers no picture of a sequential machine as a 'black box'. Hence, let us make one. Figure 5 shows a sequential machine involving two Trivial Machines, whose operations follow the relations in tables such as Moore's. Figures 6, 7, and 8 show the machine's presumed three-step input/output cycle.

Sequential machines have broad importance. They are cases of what von Foerster [14-15] later calls *Non-Trivial Machines*. Like Moore [12], von Foerster provides an example in the form of two tables ([15], p. 11). The tables describe the output **y** and the next state $\mathbf{z}'$ in terms of the input **x**, but for only two possible internal states, the present states **z**, dubbed **I** or **II**. Having two states characterizes the simplest Non-Trivial Machine; under

only *one* internal state, a particular input would always evoke a particular pre-determined, unchanging output, making the machine Trivial. Nonetheless, von Foerster's example inputs and outputs were the same as for his Trivial Machine: **x** = A, U, S, or T, and **y** = 0 or 1.

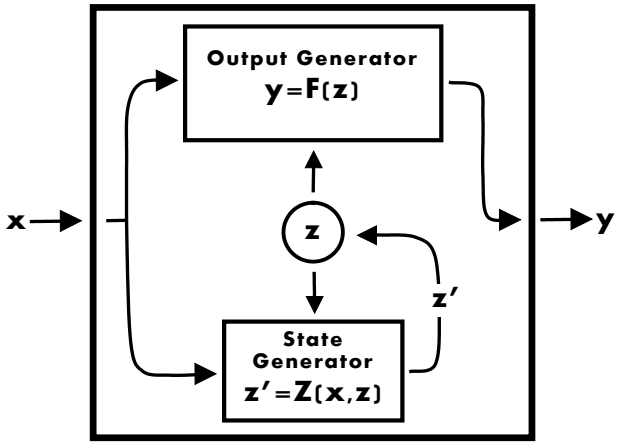

Figure 5. A Moore [12] sequential machine, depicted in a style used later by von Foerster [14-15]. The boxes and lines and the circle represent mechano-electrical parts. The lines with arrows represent the parts' operating relations, which need not occur simultaneously. The internal state **z** actively affects the Output Generator **F**, and the State Generator **Z** from which it arose. **Z** produces a new state **z′** after **y** is output by **F** when **F** is prompted by the input **x**.

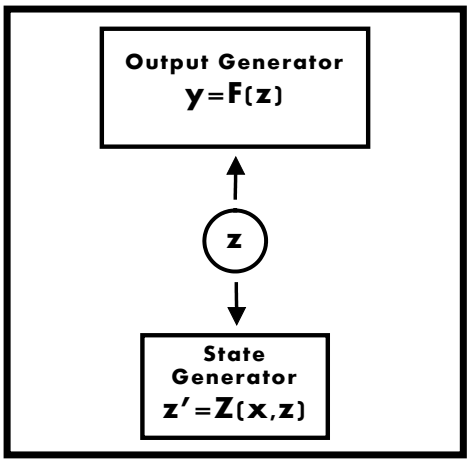

Figure 6. The operation of a Moore sequential machine, interpreted as a cycle. In Step 1, the machine awaits input while the internal state **z** actively affects the Output Generator **F** and the State Generator **Z**.

Von Foerster's data can be re-ordered, to make four new tables, two for each of **z** = **I** or **z** = **II**. Table 2 contains the four tables. Two of the tables show **y** as a function of **x**, and two of the tables show **z′** as a function of **x**. These latter pairs are the respective equivalents of the Output Generator and the State Generator of Moore's [12] sequential machine (Fig. 5). Von Foerster calls them the Driving Function and the State Function. Figure 9 schematizes von Foerster's Non-Trivial Machine.

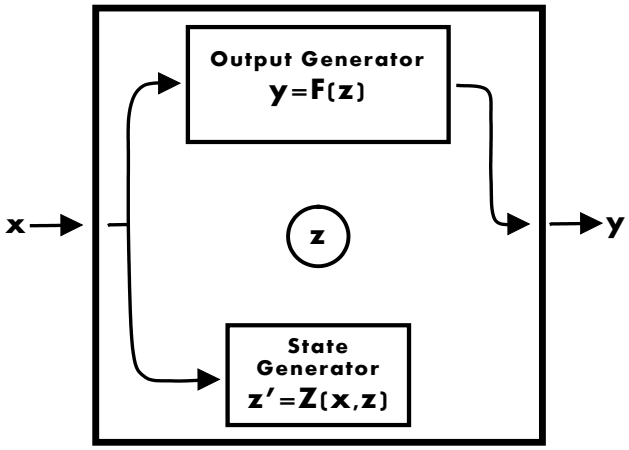

Figure 7. The imagined Step 2 of the cycle of a Moore sequential machine. The input **x** prompts an output **y** and also affects the State Generator **Z**.

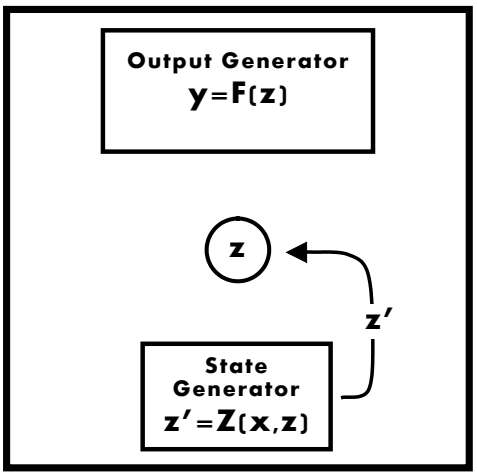

Figure 8. The imagined Step 3 of the cycle of a Moore sequential machine. After **y** is output, the State Generator generates **z′**, which replaces **z**.

Von Foerster's machine conceivably follows a three-step operational cycle like Moore's (Figs. 6, 7, and 8). But the machines in Fig. 5 and Fig. 9 profoundly differ in one detail. To Moore [12], the input **x** has no bearing on the *immediate resulting* output **y** (Fig. 5), only on its *successor* by way of the internal state. Moore's **y** is *evoked* by **x**, but is only indirectly a *function of* **x** by way of the internal state. In contrast, in von Foerster's [14-15] machine (Fig. 9), **x** directly affects **y**, *and* indirectly affects the next output by way of the internal state.

z = I	
x	**y**
A	0
U	1
S	1
T	0

z = II	
x	**y**
A	1
U	0
S	0
T	1

z = I	
x	**z′**
A	I
U	I
S	II
T	II

z = II	
x	**z′**
A	I
U	II
S	I
T	II

Table 2. Relations in von Foerster's example of a Non-Trivial Machine [15]. The two leftward tables describe the Driving Function, and the two rightward tables describe the State Function, for internal states **I** and **II**.

As an example of how the von Foerster Non-Trivial Machine works, note that when **z** = II and an input **x** = U occurs then the output **y** = 0 is evoked and the internal state changes to **z′** = II, which coincidentally is the same state as before.

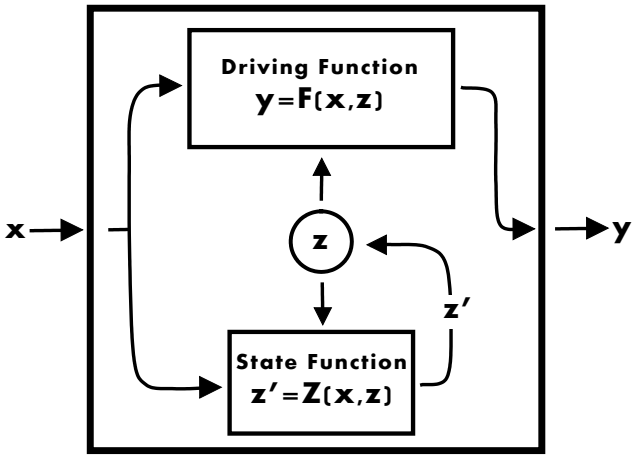

Figure 9. (After [15]) Von Foerster's "Non-Trivial Machine". It involves two Trivial Machines, the 'Driving Function' and the 'State Function', the respective operational equivalents of the Output Generator and the State Generator in Fig. 5. In Fig. 5, however, **y** is a direct function only of **z**; here, **y** is a direct function of both **z** and **x**.

10 'SYSTEMS' IN TERMS OF 'MACHINES'

To Moore, scientists belong to systems: "The experiment may not be completely isolated from the experimenter, i.e., the experimenter may be experimenting on a system of which he himself is a part" ([12], p. 133). So "The experimenter [probing a 'machine'] could be described as another sequential machine, also specified in terms of its internal states, inputs, and outputs. The output of the machine being experimented on would serve as input to the experimenter and vice versa" ([12], p. 135).

Further logic-wise (but earlier text-wise), Moore ([12], p. 132) notes that a Psychiatrist *experiments on* a patient, giving inputs and receiving outputs. Moore's 'black box' is evidently the *mind*. As Moore declares ([12], p. 132), "The black box restriction corresponds approximately to the distinction between the psychiatrist and the brain surgeon", insofar as the surgeon can alter the brain, but only the Psychiatrist can alter the mind. (Modern surgeons might disagree, but that is beside the point.)

For a sequential machine to be a mind, it would need to be capable of an enormous number of possible behaviors. Just how many behaviors, then, could possibly characterize a sequential machine? In the latter regard, von Foerster relates the case of graduate students tasked with 'interrogating' a physical Black Box ([14], p. 312), constructed by Ashby and having four distinct inputs, four distinct outputs, and four hidden distinct inner states (representing 4! = 24 permutations). This relatively simple configuration allows a truly enormous number of possible input/output combinations. Needless to say, the graduates were unable to infer the inner states.

11 HOW A FURTHER-EXTERNAL OBSERVER WOULD INTERACT WITH THE SYSTEM

Sections 7, 8, and 9 introduced the concept of the machine, as an aid to resolving a quandary. That quandary is illustrated in Figs. 3 and 4. It is the question of whether a further-external observer of the BlackBox/observer system can ignore the original, internal observer, as if knowing of that observer's presence and behavior.

To resolve the quandary, let us assume first that the core Black Box can be probed by its immediate observer without being perturbed. Suppose also that the immediate observer remains unperturbed by the output from the core Black Box. Altogether, the core BlackBox/observer system is unaltered by its internal interactions. Hence, the degree to which the immediate observer and the core Black Box understand each other will be limited only by the number of possible inputs from each to the other. *The core Black Box and its immediate observer can fully 'whiten' each other in time. They are Trivial Machines.*

This implies that if the core BlackBox/observer system is probed by a further-external observer, the latter can ignore the observer and directly interrogate the core Black Box. This direct access applies 'by induction' to all further-outward observers. This is what Glanville illustrates in 1982 [4], and is shown here as Fig. 3. The system formed by the combination of *any* observer with the core Black Box is no different than the system formed by the combination of any *other* observer with the core Black Box. The core BlackBox/observer *system* is penetrable. And it is not unique; any other, possibly further-out observer can pair with the core Black Box to form an identical system.

Consider now the alternative. Imagine that the core Black Box *cannot* be probed by its immediate observer without being perturbed, and that, similarly, the immediate observer cannot receive output from the core Black Box without changing. Box and observer are now Non-Trivial Machines. But a Non-Trivial Machine concatenated with a Non-Trivial Machine is, perforce, a Non-Trivial Machine. The core Black Box and its immediate

observer are now truly 'entangled', that is, no outside observer can tell them apart and hence *ignore* the immediate observer.

Thanks to the concept of machines, we can now comprehend the difference between the portrayal of the Black Box and its observer in Glanville [4] and that in Glanville [7-8]. In the earlier Glanville, the Black Box and its observer are Trivial Machines; in the later Glanville, they are Non-Trivial Machines.

If the core BlackBox/observer system is a Non-Trivial Machine, then it would be perturbed if probed through input from a *further-external* observer, as in Fig. 4. Whether or not that further-external observer is himself a Non-Trivial Machine, nonetheless his concatenation with the core BlackBox/observer system is a new system which is a Non-Trivial Machine. *That* system is a Black Box to any *yet-further-external* observer, and can be perturbed by that observer. Glanville notes that "each Black Box is potentially made up of a recursion of Black Boxes (and observers)" ([7], p. 1). Figure 10 shows the recursion.

12 AT THE CORE OF ANY BLACK BOX THERE ARE TWO (OR MORE) WHITE BOXES, REQUIRED TO STAY IN

The title of Glanville's landmark paper of 1982 was "Inside every white box there are two black boxes trying to get out" [4]. Figure 2 shows this arrangement when the observer himself is a Black Box. But the arguments above suggest a new interpretation. Let us presume that the core Black Box is a Non-Trivial Machine, composed of concatenated Trivial Machines. Then, no matter how many nested layers of Black Boxes and observers might occur Russian-Doll fashion within *any* Black Box (Fig. 10), the latter Box has an utter core containing a Black Box which consists of two (or more) White Boxes, boxes that are required to stay in – observed by an observer who, if he's a Black Box himself, also consists of two (or more) White Boxes. Figure 11 schematizes the old versus new approaches to the relation of White Boxes to Black Boxes.

13 SUMMARY AND CONCLUSIONS

Sensations – and the ability to report them – characterize the mind. But no-one can directly observe their own mind, or any other. Here, we attempt to understand the mind indirectly, through the concepts of the Black Box and its observer. Ranulph Glanville proselytized these concepts after W. Ross Ashby.

Ashby's Black Box differs crucially from an engineer's or a physicist's 'black box': it is un-openable. But Glanville pushes further, taking the Black Box to be an "explanatory principle", one which nonetheless has a "mechanism". These notions well-characterize the *mind*. There are other parallels. The Black Box is interrogated by an observer, who presents stimuli to the Black Box, the inputs, and who records the stimulus-evoked responses, the outputs. This mimics Psychiatry and Psychology.

Through input/output interaction with the Black Box, the observer obtains what Glanville calls a "functional description", one which "whitens" the Black Box. Likewise, however, the Black Box may "whiten" the observer – after all, the output from one is the input to the other. Altogether, the Black Box and its observer form a self-illuminating *system*, called a 'white box', having different properties than the Black Box or the observer.

The white box is nonetheless allegedly 'black' to any further-external observer. Let us call this box the 'greater Black Box', and realize that it may be probed by a further-external observer. How far, then, do inputs from the further-external observer penetrate? Glanville illustrates them going straight through the conceptual boundary of the greater Black Box and right up to the original, core Black Box itself, without interacting with that Black Box's immediate observer – as if the latter's presence and behavior were already 'visible'. Regardless, Glanville later shows the further-external observer *not* penetrating the greater Black Box. This significant change remains unresolved.

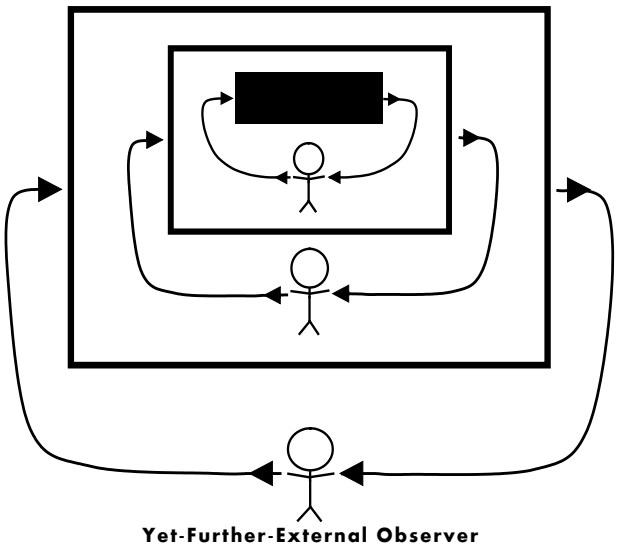

Figure 10. From the viewpoint of a *yet-further-external* observer, the greater system shown in Fig. 4 is a Black Box; and so on, with each further-outwards Black Box having its own immediate observer.

Here, a resolution is offered. It employs the concept of the 'machine', as used by von Foerster: namely, an abstract entity produced by a mechanoelectrical basis. Such an abstract entity is the Glanville Black Box. 'Machines'-wise, von Foerster follows Turing, E.F. Moore, and Ashby in recognizing archetypes that he calls the Trivial Machine and the Non-Trivial Machine. A Trivial Machine is characterized by a particular input always evoking a particular output, altogether providing an unchanging set of *relations* that constitutes the actual machine. But a mechanoelectrical basis can have internal configurations, denoted 'states'; and so, in principle, can Black Boxes. States can change in response to input or output, such that a particular input need not result in the same output later. This characterizes Non-Trivial Machines, which can have an enormous range of such 'behaviors'. The mind, too, has numerous 'states', allowing a broad range of behaviors, and those 'states' can change or increase in number through learning. For example, our response to a stimulus (such as an event, or a question) can differ from our previous response, and in unexpected ways. The mind is a Non-Trivial Machine, an abstract entity having a mechanoelectrical basis; it is a Black Box, an explanatory principle.

We now ask whether the observer of any Black Box can provide input, and record output, without changing the Box's possible output to the next input. Likewise, we must ask whether the Black Box leaves its own observer unperturbed. Consider answering "Yes" to both questions. If so, the observer's knowledge of the Black Box, and the Black Box's knowledge of its observer, will be limited only by the variety of the inputs from each to the other. *The Black Box and its observer are now Trivial Machines*. They will mutually discover this in time, as they 'whiten' each other through input and output.

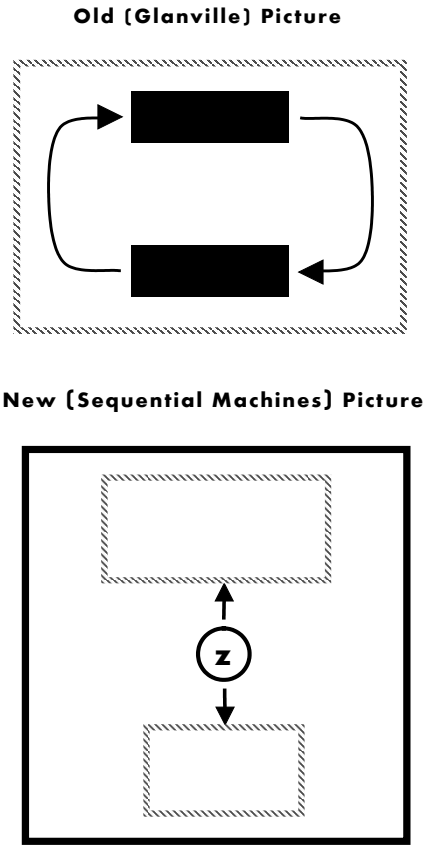

Figure 11. White Boxes versus Black Boxes. (Upper) The Glanville [4] view that "Inside every White Box there are two Black Boxes trying to get out". (Lower) The view that at the utter core of any Black Box there are two (or more) White Boxes, required to stay in.

Now consider the contrary. That is, imagine a Non-Trivial Machine that is the Black Box. Imagine another Non-Trivial Machine, that is the observer. The BlackBox/observer duo continually alter each other as each receives inputs and produces outputs. Altogether, then, the BlackBox/observer duo form a new Non-Trivial Machine. In this case, no further-external observer can differentiate the Black Box from its observer; the internal observer cannot be *recognized*, and hence it cannot be bypassed. The BlackBox/observer duo is now truly a *system*.

This system, the greater Black Box, a Non-Trivial Machine, will change when probed by the *further-external* observer. That is, the further-external observer and the greater Black Box altogether constitute a *new* system, which itself is a Black Box and a Non-Trivial Machine. Likewise, this new system changes when probed by a *yet-further-external* observer. This may continue, in a recursion of Black Boxes and their observers.

Glanville's seminal paper (1982) was titled "Inside every White Box there are two Black Boxes trying to get out". Instead, we can say that at the utter core of any Black Box there are two (or more) White Boxes, required to stay in. Those two or more White Boxes may be considered the ultimate source of the mind.

REFERENCES

[1] L. Nizami. I, NEURON: the Neuron *as* the Collective. *Kybernetes*, 46:1508-1526 (2017).

[2] L. Nizami. Too Resilient for Anyone's Good: 'Infant Psychophysics' Viewed through Second-order Cybernetics, Part 1 (Background and Problems). Early online posting, *Kybernetes* (2018).

[3] L. Nizami. Too Resilient for Anyone's Good: 'Infant Psychophysics' Viewed through Second-order Cybernetics, Part 2 (Re-Interpretation). Early online posting, *Kybernetes* (2018).

[4] R. Glanville. Inside Every White Box there are Two Black Boxes Trying to Get Out. *Behavioral Science*, 27:1-11 (1982).

[5] R. Glanville. Behind the Curtain. In: R. Ascott (Ed.), *Procs. First Conf. on Consciousness Reframed*, UCWN (University of Wales College Newport), 5 pages, not numbered (1997).

[6] R. Glanville. A (Cybernetic) Musing: Ashby and the Black Box. *Cybernetics and Human Knowing*, 14:189-196 (2007).

[7] R. Glanville. Darkening the Black Box. Abstract. In: *Procs. 13th World Multi-Conference on Systemics, Cybernetics and Informatics*, Orlando, FL, International Institute of Informatics and Systemics (2009).

[8] R. Glanville. Black Boxes. *Cybernetics and Human Knowing*, 16:153-167 (2009).

[9] M. Ramage and K. Shipp. *Systems Thinkers*. Springer, New York, NY (2009).

[10] W. Ross Ashby. *An Introduction to Cybernetics*. Fourth Edition; Chapman & Hall Ltd., London, UK (1961).

[11] L. Nizami. Homunculus Strides Again: Why 'Information Transmitted' in Neuroscience Tells Us Nothing. *Kybernetes*, 44:1358-1370 (2015).

[12] E.F. Moore. Gedanken-Experiments on Sequential Machines. In: C.E. Shannon & J. McCarthy (Eds.), *Automata Studies* (Annals of Mathematics Studies Number 34). Princeton University Press, Princeton, NJ, pp. 129-153 (1956).

[13] https://www.aaschool.ac.uk/PUBLIC/NEWSNOTICES/obituaries.php?page=4, http://www.isce.edu/Glanville.html

[14] H. von Foerster. *Understanding Understanding: Essays on Cybernetics and Cognition*. Springer-Verlag, New York, NY (2003).

[15] H. von Foerster. Principles of Self-Organization – in a Socio-Managerial Context. In: H. Ulrich & G.J.B. Probst (Eds.), *Self-Organization and Management of Social Systems*. Springer Series in Synergetics, vol 26. Springer, Heidelberg, Germany, pp. 2-24 (1984).

[16] A.M. Turing. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42:230-265 (1937).

[17] L. Nizami *Reductionism ad Absurdum*: Attneave and Dennett Cannot Reduce Homunculus (and Hence the Mind). *Kybernetes*, 47:163-185 (2018).

From tools to social agents

Anna Strasser[1]

Abstract. Up to now, our understanding of sociality is neatly tied to living beings. However, recent developments in Artificial Intelligence make it conceivable that we will entertain interactions with artificial systems which have social features in the near future. By suggesting a minimal approach to a conceptual framework of socio-cognitive abilities, this paper presents a strategy of how social interactions between humans and artificial agents can be captured. Taking joint actions as a paradigmatic example, minimal necessary conditions for artificial agents are elaborated. To this end, it is first argued that multiple realizations of socio-cognitive abilities can lead to asymmetric cases of joint actions. In a second step, it is discussed how artificial agents can meet minimal conditions of agency and coordination in order to qualify as social agents in joint actions.

1 INTRODUCTION

Soon we will share a large part of our social lives with various kinds of artificial systems. Besides using them as tools in order to interact with other human agents or to retrieve information, it is conceivable that we will as well entertain social interactions with artificial agents. Even though most of those interactions can sufficiently be described as tool use, it is worthwhile to investigate whether artificial agents may possess socio-cognitive abilities which enable them to overtake the role of a social agent and thereby constitute a new range of social interactions.

For example, think about future interactions with care-robots and conversational machines (chatbots). Imagine an older person spending most of her time with a care-robot – not only using this robot as an assistant but also to communicate and to satisfy her social needs. I claim that categorizing such interactions as mere tool use will neglect important social aspects. Having such interactions in mind, it is an urgent challenge to explore the potential role of artificial agents in the realm of social cognition and describe circumstances in which artificial agents could be considered as social agents and not as mere tools.

However, our current conceptual framework concerning social agents cannot account for artificial agents as social agents. To overcome this restriction, we have two options: We can either propose an extension of the concept of tools, claiming that there are more complex tools which have some social features. Alternatively, we can contemplate an extension of our conception of social agents. Assuming that there are interactions conceivable for which it is at least questionable whether we can classify them as tool use, I will pursue the latter.

2 RESTRICTIVE UNDERSTANDING OF SOCIALITY

Up to now, our understanding of sociality is neatly tied to living beings. Social cognition is treated as a distinguishing feature of living beings. Research about social cognition includes topics such as social knowledge, social structure, group behavior, social influence, memory for social information, and attribution of motives [1]. All this is explored in humans and several species of the animal kingdom.

Focusing on the practice of ascribing socio-cognitive abilities, one can object that there are two modes of ascriptions: an ‘as if’ mode and a justified mode of ascription. The ‘as if’ mode has an explanatory role, it helps us to make sense of the world, but it remains neutral about the question of what socio-cognitive abilities objects really have. For instance, a famous experiment by Heider and Simmel [2] illustrates how participants attribute social properties while describing simply moving geometrical forms. Although it is helpful and enlightening to characterize perceptual input through a social narrative and not through a technical description of geometric forms, it is of course not justified to claim (and no one does) that these objects actually have social features. Along the same lines, Daniel Dennett [3] describes how we apply the intentional stance to non-living beings.

Turning to standard philosophical notions which characterize socio-cognitive abilities as if they were unique to sophisticated adult human beings, we are even confronted with more restrictive notions. For example, the notion specifying individual agency [4] requires demanding conditions such as consciousness, the ability to generate goals, the ability for free choice, having propositional attitudes, mastery of language, and intentionality. Likewise, the notion of joint action as introduced by Michael Bratman [5] presupposes cognitively demanding conditions. Without having the ability to entertain shared intentions, agents do not qualify for joint actions. Thus, having shared intentions requires not only having an intention but also the ability to entertain a specific belief state, namely a relation of interdependence and mutual responsiveness which in turn presupposes common knowledge. According to Bratman, only participants who are able to coordinate and build up explicit relations of commitment qualify as proper social agents in joint actions.

However, such notions account for full-fledged, ideal cases, and cannot capture other forms of realizations with less demanding or simply different requirements. Research in developmental psychology, as well as in animal cognition, indicates that there are multiple realizations of socio-cognitive abilities. Moreover, even with respect to adult humans, one can observe that such ideal cases occur less often than expected. For

[1] Independent Researcher, Berlin, Germany.
Email: annakatharinastrasser@gmail.com

instance, under time pressure otherwise sophisticated adults fall back on less demanding realizations. Likewise, developing expertise in a skill is often based on the automatization of formerly sophisticated processes.

Even though research indicates that there are multiple realizations of socio-cognitive abilities in various types of agents such as infants and non-human animals [6, 7, 8], non-living beings still are in principle excluded from having social capacities. Therefore, the aim of considering non-living artificial agents as social agents presents a revolutionary challenge. In order to investigate how to account for sociality with respect to artificial agents, one has to explore how to overcome the restrictive nature of our current understanding of sociality.

Once we have a conceptual framework which is able to capture socio-cognitive abilities of artificial systems, further questions concerning the consequences of potential social interactions will arise. And last but not least, analyzing potential consequences, one can evaluate whether developing artificial social agents is a desirable goal at all.

Inspired by the strategy of so-called minimal approaches [9, 10], which offer notions for socio-cognitive abilities of infants and non-human animals, this paper discusses potential minimal necessary conditions with respect to artificial agents. The aim is to elaborate under which circumstances interactions with artificial agents qualify as social even though we know that artificial agents are not living beings. When artificial systems prove to be capable of displaying socio-cognitive abilities, this will constitute a new category of social interaction that still is reasonably similar to those we observe among humans or other living beings.

Recent studies in social neuroscience already show that interactions with artificial agents (avatars) are at least somehow comparable to interactions among humans. On the one side, social scientists study avatars as a way of understanding people [11]. Such studies investigate interactions between humans while they are embodied in avatars. Thereby special features of interacting with virtual representations are explored. On the other side, artificial agents are used in experimental designs in which participants are tricked in the sense that they believe that the interaction partner is a human counterpart while they are actually interacting with an artificial agent. If interactions with artificial systems would not have any similarities with human-human interactions, we could not use them to explore human behavior. However, it is important to note that this paper is not about working out to what extent people might be tricked by an artificial agent and attribute social characteristics in an ‘as if’ manner. Just taking an intentional stance [3] does not yet justify an attribution of socio-cognitive abilities as such. The aim here is to investigate whether artificial systems can actually have socio-cognitive characteristics. To explore this theoretical possibility of finding socio-cognitive abilities in artificial agents we have to be cautious not to mix ‘as if’ ascriptions with justified ascriptions. The question as to whether we are justified in ascribing socio-cognitive abilities to artificial systems is based on the assumption that the increasing experience of interacting with artificial agents is likely to alter our understanding of *social agents* radically.

3 CONSEQUENCES OF ARTIFICIAL SOCIAL AGENTS

The possibility that artificial agents might qualify as proper social agents in interactions with humans will raise new ethical and juristic questions. As soon as an artificial agent is understood as a social agent, we need to pose questions about the duties and rights this artificial agent deserves as an interaction partner.

Regardless of whether an artificial agent is considered as social agent or as tool, it is common sense that artificial agents should not harm other living beings. However, as soon as we treat artificial agents as social agents (and not ‘as if’ they were social agents), this might involve ascribing rights to them. Since our self-understanding of fairness and justice is based on how we treat other social agents, it will be important to develop social norms of how to treat artificial social agents. Already in a transitional phase, when artificial systems are not yet proper social agents, our interactions with them can influence our behavior towards other living social agents. Imagine a person using an artificial agent which is strikingly similar to a human in order to satisfy some felt needs, needs fulfillment which most people would find shameful, if not downright criminal, if it were to be acted out with another human agent. How can we preclude that this behavior with an artificial agent might not make it more probable that the person would end up crossing the line between fantasy and reality in the public world?

Furthermore, regarding the outcomes of joint actions in which artificial social agents are involved, we have to face new questions of responsibilities. For example, regarding responsibilities of autonomous driving systems, we might ask whether a person who is using an autonomous vehicle while having no way of taking over control, still should be held responsible for possible accidents [12]. The more autonomous artificial systems become, the more pressing becomes the question whether simply the producers and the users are alone accountable for the outcome of the actions of those systems.

Where previous revolutions have dramatically changed our environments, this one has the potential to change our understanding of sociality and may lead to new social norms.

4 SOCIO-COGNITIVE ABILITIES OF ARTIFICIAL AGENTS

Accounting for socio-cognitive abilities of artificial agents, we have to assume that there are further multiple realizations that are not covered by our current conceptual framework. Recent research has already shown that there are certain multiple realizations of socio-cognitive abilities. For instance, we have data about how non-human animals and also very young infants are able to demonstrate social competences which presumably are based on socio-cognitive abilities. Even though there are controversial debates about how exactly such competences are realized, it is obvious that the requirements of full-fledged, ideal cases are not fulfilled in such instances.

For example, it is common sense to assume multiple realizations regarding the ability to anticipate the behavior of others. Traditional conceptions of mindreading require the mastery of language as a necessary condition. But, the competence to anticipate the behavior of other agents has also been observed in populations where we cannot assume mastery of language is operating. Admittedly, interpretations of how this competence is realized in non-verbal populations are controversial. Some positions claim that those competences are best explained by the application of behavioral rules [13, 14, 15] and thereby deny mentality to such realizations. Others argue for genuine mindreading abilities [16, 17]. This debate is far from

being decided; up to now, neither the behavioral nor the mentalistic interpretation can yet exclude the other. Consequently, the question as to whether infants or non-human animals possess the socio-cognitive ability of mindreading is still an open question. Therefore, it seems feasible at least to pose the question as to whether artificial systems can display socio-cognitive abilities.

Starting with the assumption that our understanding of socio-cognitive abilities is too restrictive, the exploration of minimal conditions for socio-cognitive phenomena concerning artificial systems can suggest an extension of our current conceptual framework for attributing socio-cognitive abilities.

5 CONDITIONS FOR JOINT ACTIONS

Joint actions constitute an interesting subset of social interactions, a subset in which people cooperate and do things together in order to reach a common goal. Taking joint actions as a paradigmatic example, this paper discusses necessary minimal conditions which qualify artificial agents as proper participants in a joint action, and specifically as social agents. A minimal notion of joint action can distinguish tool use from joint actions and thereby enable a finer-grained description of, for example, human-computer interactions.

Although the philosophical debate about joint actions is rather controversial, one can summarize some important requirements which can be seen as necessary. Disagreements start when it comes to questions of sufficiency. An event can only qualify as a joint action if it results from the input of multiple agents. That means the effect of this event can be described as a common outcome of what several agents did, and whereby the *individual agencies* are intentional under some description [4]. To distinguish mere plural activities from joint actions, we must require that both agents *aim* at bringing about the same effect. As a consequence, some sort of *coordination* is required in addition. And it is further claimed that this coordination is achieved through special psychological mechanisms. However, the question as to whether these mechanisms can be based on shared goals (weak sense of joint action), or whether these mechanisms have to include shared intentions (strong sense of joint action) to ensure that not only the same goal is achieved but also that this goal is jointly aimed at, is still under debate.

Both strategies are problematic. The weak sense of joint action captures cases in which individual agents treat each other as social tools, whereas the stronger sense requires overly demanding conditions which are, for example, not fulfilled by young children. Inspired by Pacherie's notion of 'intention lite' [18], I assume that there are middle cases which can exclude the social tool cases and at the same time refer to less demanding conditions.

To approach a solution, I first argue that there are *asymmetric cases of joint actions* in which the distribution of abilities is not equal among the participants. Uncontroversial cases of asymmetric joint actions are, for example, mother-child interactions. Despite the fact that infants do not fulfill the full-fledged sophisticated conditions of a strong sense of joint action, they are regarded as social agents in joint actions. That means they are able to act jointly with an adult participant while their fulfilled conditions differ from those of the adult. Consequently, it is sufficient to require less demanding conditions from one participant of a joint action. The performance of the participants in an asymmetric joint action can be based on multiple realizations. Consequently, artificial agents do not have to fulfill the very same conditions required of human adults.

Since any notion of joint actions describes a plural activity, one has to presuppose the ability to act. Already at this point, we need an alternative notion of agency different from that which standard philosophical positions [4] offer. In order to capture the notion of agency operative in artificial systems, we need a notion which does not rely on features we find only in biological systems. I have developed elsewhere a minimal notion of agency which does not rely on biological constraints [19, 20], and makes sure that artificial agents are worthy of being considered as potential actors. If artificial agents are not able to act in the appropriate sense, any further questions as to whether they might qualify as acting *jointly* would, of course, be a waste of time. For the sake of argument, I presuppose here that artificial systems can qualify as minimal actors. In line with the conception of asymmetric joint actions, a joint action performed by a mixed group of humans and artificial agents can then be seen as a combination of two types of agency.

The ability to coordinate will be at center stage in this investigation because coordination plays a crucial role for constituting the social dimension of joint actions. Regardless of whether one assumes shared goals or shared intentions, successful coordination in social interactions presupposes social competence. Agents must have some sort of an understanding of the other agents which makes it possible to anticipate the other's behavior and to rely on the other's willingness to take over its part. Consequently, mindreading and commitment are seen as important factors for ensuring the social competence needed for coordination which is necessary in joint actions.

6 ANTICIPATION – MINDREADING

It is common sense that a major function of social cognition consists in abilities to encode, store, retrieve, and process social information about conspecifics, as well as across species, in order to understand others. One important aspect, namely the ability to anticipate the behavior of other agents, plays an important role in many social interactions. Being able to act jointly we have to be able to anticipate what the other agent will do next. In the humanities and natural sciences, this aspect of social competence is discussed under the label of 'mindreading' or 'Theory of Mind' [21].

If artificial agents qualify as social agents in a joint action, we have to expect mindreading abilities from them. As we have already seen with respect to the notion of agency, standard notions tend to be rather restrictive and demanding. The same is true for mindreading. Many conceptions of mindreading are tailored to adult humans and refer to a full-fledged form of mindreading requiring a mastery of language, as well as cognitively demanding abilities such as meta-representations.

Assuming multiple realizations and building upon minimal approaches, one can elaborate minimal necessary conditions of mindreading. In this paper, I am arguing that the less demanding conditions for *minimal mindreading* [9] provide an attractive alternative to capture the mindreading abilities of artificial agents.

In contrast to full-fledged mindreading, this minimal approach specifies minimal presuppositions for mindreading. Instead of requiring a wide range of complex mental states, Butterfill and Apperly [9] specify two mental states, namely encounterings and registrations. Roughly speaking, one may characterize encounterings as a kind of simple perception, whereas

registrations could be described as a rudimentary form of believing. A minimal mindreader infers from observable cues to the mental state of encountering. With respect to the last observed encountering, the minimal mindreader then ascribes a further mental state (registration) to the other agent. Finally, she applies a minimal theory of mind – which consists in the knowledge that goal-directed actions rely on registrations – to anticipate the behavior of the other. Minimal mindreaders can in a limited but useful range of situations track others' perceptions and beliefs without representing perceptions and beliefs as such, but instead representing encounterings and registrations. Minimal mindreading is regarded as implicit, nonverbal, automatic, and is based on unconscious reasoning.

Research in artificial intelligence has already demonstrated that artificial agents can model mental states of human beings with respect to the perspective of a human counterpart [22]. This shows that artificial agents, in principle, are able to infer from their perception of the physical world to what a human counterpart can see or cannot see in terms of an object and are capable of inferring that this perspective will guide future actions of the human. That means some cases of mindreading can be achieved by artificial agents.

At this point, one can object that we are neglecting the genuine social aspect of mindreading. Admittedly, many examples in the mindreading debate tend to relate to mental states such as knowing and perceiving. Desires and emotions are not yet at the foreground of these debates. Focusing on the genuine social aspect, one can conclude that qualifying as a mindreader should include the ability to process social information. For instance, we do not only have to notice that another agent is noticing something that is relevant for the joint action, but we should also recognize whether the other agent is desiring something or is afraid of something. This presents a special challenge for artificial agents. Taking into account that human anticipatory systems fairly seamlessly include social and emotional aspects, we have to explore whether artificial systems are able to process such data as well. To anticipate future actions of other agents, it is not only relevant to consider their mental, but also their emotional states.

Turning to emotional data, actual research on social robotics is highly relevant, specifically in relation to the development of robots which are designed to enter the space of human social interaction. For example, research pertaining to conversational agents aims to develop artificial agents from mere tools into human-like partners [23, 24]. Since the processing of social data plays an important role in social interactions, social relationships between artificial agents and humans presuppose that the artificial agents interpret the social cues presented by their interacting partners. And they should also be able to send social cues in order to make their ‚minds' visible.

Much research is now focusing on social cues such as gestures [25] and emotional expression [26, 24]. For example, ARIAs (Artificial Retrieval of Information Assistants) [27] are able to handle multimodal social interactions. They can maintain a conversation with a human agent and, indeed, they react adequately to verbal and nonverbal behavior. Even though results in social robotics may not apply to an unlimited range of situations, this shows that there are ways for artificial agents to process social data.

The above considerations indicate that artificial agents, in principle, are able to process social data and make use of it to anticipate the behavior of their interaction partners. Further developments in social robotics will probably also make it easier for the human counterpart to anticipate the behavior of the artificial agent.

However, according to traditional philosophical notions of mindreading, mere processing of emotional data is not taken as sufficient. In addition, having emotional and mental states is required. Assuming that mental or emotional states are exclusively found in living beings, our question as to whether artificial agents can be social interaction partners in a joint action turns into the question as to whether having mental and emotional states is a necessary requirement for realizing socio-cognitive abilities such as mindreading. One might argue that future AI systems might someday have mental and emotional states. But up until now, it does not look like as if this is to be expected in the near future. Therefore, the crucial question is whether we can ensure that we are not losing the sociality aspect even if we sacrifice mental and emotional states.

So far, the notion of minimal mindreading [9] is a promising starting point to characterize mindreading abilities of artificial agents. As we have seen, this notion questions the necessity of overly demanding cognitive resources, such as the ability to represent a full range of complex mental states and a mastery of language. And most importantly for artificial agents, the ability for minimal mindreading need not be based on conscious reasoning. Nevertheless, up to now, this notion has been only applied to living beings, only accounting for automatic mindreading in human adults, infants, and non-human animals. Even though this notion does not require conscious reasoning from a mindreading agent, future work will have to deliver further adjustments before it can be applied to artificial systems.

In sum, one can argue that, in principle, artificial systems are able to process social and mental data and use it with a Theory of Mind to anticipate the behavior of human agents and thereby qualify as mindreaders. In a transition phase, it is likely that this works only in a very limited range of situations and it might be a special feature of asymmetric joint actions that they always only constitute a limited subset of joint actions.

7 COMMITMENT

Another aspect of the required social competence enabling successful coordination in a joint action can be described as the ability to be committed to a joint action. To explore commitments with respect to artificial agents, the recently developed notion of a minimal sense of commitment [10] presents a good starting point.

Commitments are relations between agents and an action which provide the security human social agents need to rely on each other. Additionally, commitments support the success of mindreading, since the behavior of agents who are sticking to their commitments is far easier to be predicted. In sum, one can claim that commitments function as the 'social glue' for much of what counts as social interactions.

Standard philosophical conceptions [28, 29, 30] characterize commitments as a relation between two or more agents and a specific action: An agent is committed to performing a specific action if she has assured her commitment and the other agent has acknowledged this. One component of a commitment is based on the motivation of one agent to contribute a specific action to a joint action; the other component is based on the corresponding expectation of the other agent that the counterpart will contribute

to the joint action. Additionally, it can be claimed that this requires explicit acknowledgment and common knowledge. Standard conceptions of commitments rely on explicit utterances and are interpersonal since they describe a reciprocal relation between (at least) two agents. This can be contrasted with self-commitments which require just one agent.

Analyzing the possible classes of interpersonal commitments, it becomes obvious that standard conceptions neglect other potential cases. For example, not all interpersonal commitments require necessarily explicit assurances and acknowledgments. We experience implicit commitments in everyday life situations when agents feel and act committed even though no commitment was explicitly acknowledged [31]. Research in developmental psychology indicates cases of implicit commitments by showing that young children are capable of engaging in joint actions which rely on an interpersonal commitment without an explicit acknowledgment [32]. Therefore, it seems uncontroversial to claim that commitments can also be realized in an implicit way.

Coming back to the notion of a minimal sense of commitment [10], we have a minimal approach to interpersonal commitments within which implicit commitments are also captured. It is of special interest with respect to the aim of this paper that this minimal approach additionally illuminates other neglected minimal forms of interpersonal commitments. Michael and colleges [10] argue that components of a standard commitment, namely the expectation or the motivation, can be disassociated. Consequently, they claim that a single occurrence of just one component can be treated as a sufficient condition for a minimal sense of commitment. Presupposing that there is a goal of a potential joint action desired by one agent for which an external contribution of another agent is crucial, a minimal sense of commitment is already constituted if either one of the agents has a certain motivation, the other has a specific expectation, or both entertain the corresponding mental states.

In the standard cases, expectations are justified by the motivation of the other agent, whereas in minimal cases the expectation of one participant can be sufficient. Applying this to asymmetric joint actions, a minimal sense of commitment realized by one participant (e.g., the human) can be sufficient. Assuming that artificial agents neither have emotional nor mental states displaying a minimal sense of commitment presents a real challenge for attributing commitments to them. Future work will investigate whether artificial agents can display functionally equivalent states according to which it becomes reasonable to ascribe a minimal sense of commitment to them. However, with respect to asymmetric joint actions, it is, for the most minimal case, sufficient if only human counterparts entertain a minimal sense of commitment.

8 CONCLUSION

Presupposing that artificial agents become increasingly prevalent in human social life, it is important to examine whether we are justified in ascribing socio-cognitive abilities to them, and go on from there to consider artificial agents as social agents.

Starting with an examination of current and rather restrictive conceptions of sociality, this paper explored minimal necessary conditions enabling artificial agents to enter the realm of social cognition. One question was whether it can be a function of social cognition to encode, store, retrieve, and process social information not only concerning conspecifics or other species but also regarding artificial agents. Another question was whether artificial agents can have social cognition to encode, store, retrieve, and process social information concerning human beings.

Building upon multiple realizations of socio-cognitive abilities, I argued that there are asymmetric cases of joint actions in which the distribution of abilities is not equal among the participants. Therefore, artificial systems could take advantage of this asymmetry, which applies in some human cases of joint actions, so that, as has been argued, they do not have to fulfil the same – and idealized – conditions that are normally assumed to be fulfilled by living beings.

To this end, I suggested a minimal approach to joint actions when characterizing a joint action between artificial and human agents. I suggested easing the standard requirements for joint actions, which are based on demanding conceptions of agency and coordination. In a first step, I suggested replacing the demanding notion of agency with a minimal notion of agency according to which artificial systems can be seen as, at least, potential actors. In a second step, presuppositions of successful coordination in joint actions were analyzed. The social competence to anticipate the behavior of other agents (mindreading) and to rely on their willingness to take over their part (commitment) were at the center of this investigation.

Developing minimal conditions for the requested social competence, I questioned whether having mental or emotional states is a necessary condition. Not requiring mental or emotional states is crucial for maintaining that proposed minimal conditions still can ensure that we are not losing the sociality aspect, on which we are focusing when we discuss whether artificial agents qualify as social agents.

With respect to mindreading, a possible obstacle for artificial agents may be the ability to process and interpret social data such as gestures, facial expressions, and gaze following. However, developments in social robotics demonstrate that processing such social data is at least not impossible. It may not yet be sufficient to cover all sorts of social interactions, but it can cover a subset of social interactions. If having mental and emotional states is not a necessary requirement for successful processing of social data, a more completely developed notion of minimal mindreading [9] has the potential to capture the notion of social competence in artificial systems.

Focusing on the question as to under which circumstances a sense of commitment may arise in such interactions, considerations about the recently developed notion of a minimal sense of commitment [10] indicate how commitments can play a role in joint actions with mixed groups of artificial and human agents.

In sum, this sketch of a variety of minimal approaches describes joint actions of mixed groups of humans and artificial agents as a combination of two different sets of requirements. Whether interactions between two artificial agents may have social features will be a topic of future research.

In limited situations, we might even now claim that, for example, conversational machines are able to coordinate their speech acts to the speech acts of their dialogue partner, and thereby meet an important condition for joint action. Whatever future research will bring, with a conceptual framework that clarifies requirements for social agents, we can better characterize, understand and regulate potential social interactions with artificial agents.

REFERENCES

[1] U. Frith and S.-J. Blakemore. Social Cognition. In: *Cognitive Systems. Information Processing Meets Brain Science*. R. Morris, L. Tarassenko, M. Kenward (Eds.). Elsevier Academic Press (2006).

[2] F. Heider and Simmel, M. An experimental study of apparent behavior. *The American Journal of Psychology*, 57, 243–259 (1944).

[3] D. Dennett. The Intentional Stance. MIT Press. (1987).

[4] D. Davidson. *Essays on actions and events*. Oxford: Oxford University Press. (1980).

[5] M. Bratman. *Shared Agency: A Planning Theory of Acting Together*. Oxford: Oxford University Press. (2014).

[6] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral Brain Sciences, 1*, 515–526 (1978).

[7] C. Heyes. False belief in infancy: a fresh look. *Developmental Science, 17*(5), 647–659 (2014).

[8] C. Heyes. Animal mindreading: what's the problem? *Psychonomic Bulletin & Review, 22*(2), 313–327 (2015).

[9] S. Butterfill and I. Apperly. How to construct a minimal theory of mind. *Mind and Language*, 28(5), 606–637 (2013).

[10] J. Michael, N. Sebanz and G. Knoblich. The Sense of Commitment: A Minimal Approach. *Frontiers in Psychology*, 6, 1968 (2016).

[11] J. Scarborough and J. Bailenson. Avatar Psychology. In: *The Oxford Handbook of Virtuality*. Oxford University Press. (2014).

[12] A. Hevelke and J. Nida-Rümelin. Responsibility for Crashes of Autonomous Vehicles: An Ethical Analysis. *J. Sci Eng Ethics*, 21: 619 (2015).

[13] C. Heyes. Theory of Mind in Nonhuman Primates. *Behavioral and Brain Sciences*, 21(01) (1998).

[14] D. Penn and D. Povinelli. On the Lack of Evidence that Non-human Animals Possess Anything Remotely Resembling a 'Theory of Mind'. *Philosophical Transactions of the Royal Society B* 362 (1480), 731–744 (2007).

[15] A. Whiten. Humans Are Not Alone in Computing How Others See the World. *Animal Behaviour* 86(2), 213–221 (2013).

[16] L. Fletcher and P. Carruthers. Behavior-Reading versus Mentalizing in Animals. In: *Agency and Joint Attention*. J. Metcalfe & H. Terrace (eds.). Oxford: Oxford University Press, 82–99 (2013).

[17] M. Halina. There Is No Special Problem of Mindreading in Nonhuman Animals. *Philosophy of Science* 82(3), 473–490 (2015).

[18] E. Pacherie. Intentional joint agency: Shared intention lite. *Synthese*, 190(10):1817–1839 (2013).

[19] A. Strasser, *Kognition künstlicher Systeme*. Frankfurt: Ontos-Verlag. (2005).

[20] A. Strasser. Can artificial systems be part of a collective action? In: *Collective Agency and Cooperation in Natural and Artificial Systems. Explanation, Implementation and Simulation*. C. Misselhorn (ed.) Philosophical Studies Series, Vol. 122. Springer. (2015).

[21] J. Fodor. Theory of the Child's Theory of Mind. *Cognition*, 44(3), 283–296 (1992).

[22] J. Gray and C. Breazeal. Manipulating Mental States Through Physical Action – A Self-as-Simulator Approach to Choosing Physical Actions Based on Mental State Outcomes. *International Journal of Social Robotics*, 6(3), 315–327 (2014).

[23] N. Mattar and I. Wachsmuth. Small talk is more than chit-chat: Exploiting structures of casual conversations for a virtual agent. In: *KI 2012: Advances in artificial intelligence, Lecture notes in computer science*, vol. 7526, 119–130. Berlin: Springer. (2012).

[24] C. Becker, I. Wachsmuth. Modeling primary and secondary emotions for a believable communication agent. In: *Proceedings of the 1st Workshop on Emotion and Computing in conjunction with the 29th Annual German Conference on Artificial Intelligence (KI2006)*, Bremen, pp. 31–34. (2006).

[25] S. Kang, J. Gratch, C. Sidner, R. Artstein, L. Huang, L.P. Morency. Towards building a virtual counselor: modeling nonverbal behavior during intimate self-disclosure. In: *Eleventh International Conference on Autonomous Agents and Multiagent Systems*, Valencia, Spain. (2012).

[26] P. Petta, Pelachaud, C., Cowie, R. (eds.) *Emotion-Oriented Systems: The Humaine Handbook*. Heidelberg: Springer. (2011).

[27] T. Baur, G. Mehlmann, I. Damian, P. Gebhard, F. Lingenfelser, J. Wagner, B. Lugrin, E. André. Context-aware automated analysis and annotation of social human-agent interactions. *ACM Trans. Interact. Intell. Syst.*, 5, 2 (2015).

[28] J. Austin, *How to do Things with Words*. Cambridge, MA: Harvard University Press. (1962).

[29] J. Searle, *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press. (1969).

[30] S. Shpall. Moral and rational commitment. *Philos. Phenomenological Res.*, 88(1), 146–172 (2014).

[31] M. Gilbert. Rationality in collective action. *Philos. Soc. Sci.* 36, 3–17 (2006).

[32] F. Warneken, F. Chen and M. Tomasello. Cooperative activities in young children and chimpanzees. *Child Dev.* 77, 640–663 (2006).

The Natural Connectivity of Autonomous Systems

Steve Battle [1]

Abstract. The concept of *enaction*, or embodied cognition, described by Varela et al., aims to resolve the the dilemma of Cartesian mind-body dualism by re-casting these categories as complementary explanations of the same phenomena. This paper explores Varela's symbolic/operational abstractions of these domains and the coupling between them, as applied to autonomous, enactive robots. As observers we may describe the behaviour of a robot symbolically, in terms of its function in the world. On the other hand, the autonomous robot can also be described in purely operational terms, as a behaviour that persists solely through self-regeneration; for no other purpose than to exist. Its meaning and purpose arise almost as an irrelevant side-effect of being in the world; its structural coupling with an environment.

1 INTRODUCTION

Robots suffer from the mind-body problem insofar as their programmers routinely impose upon their mechanical bodies a mind built of code. Code brings along the baggage of the physical symbol hypothesis and representationalism, and we wonder that the robot cannot make-sense of the world.

Enactive cognition is an approach to understanding how robots can operate autonomously in the human environment. The approach is rooted in the existentialist philosophy of Martin Heidegger [10], and his ideas of the nature of human existence as being thrown into the world with all the physical needs that entails. Where a disembodied computer has no real needs, a mobile robot has very real power requirements to satisfy. An autonomous robot is thrown into this world just as we are. While this may not count as a *lived* experience, for robots are not alive, can we speak of an experience of being a robot?

Varela's *enactivism* [15] emerged alongside Maturana and Varela's Autopoiesis [12]. Where Maturana is concerned primarily with the physical self-organisation of living structures, Varela's theory of autonomous systems [14] focuses on functional self-organization alone. The latter theory is far more pertinent to AI and autonomous robotics, which cannot be considered to be alive (autopoietic). Whilst not alive, a robot must protect its own existence; its selfhood, realised as a complex of behaviours. These behaviours constitute a closed homeostatic loop with the goal of ensuring its continued survival.

The enactive approach challenges the notion of representation. Instead of seeing representations as symbols manipulated directly by brains, they should be seen instead as occurring in the interaction (structural coupling) with the world; more physical skill than computation. It is not entirely *intelligence without representation* [3], but intelligence with representation in its proper place, as symbolic, but *non-functional* explanation as seen from an observer's perspective. We cannot yet claim that robots are observers in any real sense, but we can explore lower-grade *autonomous* behaviours.

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

2 SYMBOLIC EXPLANATION

A robot is structurally coupled with its environment through sensorimotor activity. This is the interface of the robot with its environment. We can record the external behaviour of the robot; not just movement but also sensor data. We are interested in autonomous behaviours that enable robots to function semi-independently of humans.

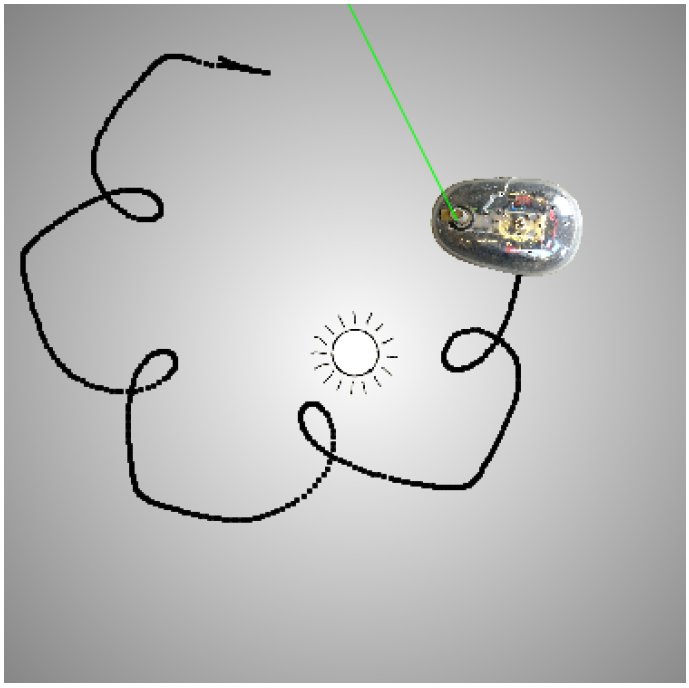

Figure 1. ELSIE simulation. Her scanning turret (direction in green) contains a single photo-cell, and her shell senses obstacles on contact.

The working example used throughout this paper is W. Grey Walter's ELSIE robot that we will refer to as *she* as befitting her name. The name is an acronym for Electro-mechanical Light-Sensing robot with Internal and External stability. She has a functional simplicity that lends itself well to examples, and yet she is also an *autonomous* robot in that her behaviour exhibits a long term viability as she is also able to periodically recharge her batteries. These *Machina speculatrix* were not digital computers, but were analogue electronic creatures. It wasn't simply that stored-program digital computers didn't

physically exist until 1948. But according to Walter, no "variety of programming endow a machine with the autonomous qualities of a true mimicry of life." Whatever the truth of this, we can explore the analogue behaviour of these creatures by simulating them in software.

In order to understand ELSIEs behaviour, a software simulation was constructed. This is not a deep simulation of her electronics, but is based on a rule-based model of her behaviour. This drives a two-dimensional kinematic simulation of the robot hardware, a three-wheel trolley with a *scanning* front wheel and light sensor. The simulated environment is populated by random obstacles and a single source of light and power. A screen-capture of the ELSIE simulation can be seen in figure 1. The output includes a trace of her recent position displaying her characteristic epicyclic trajectory.

ELSIE can sense the light-level collected by a photocell within a scanning turret containing a single photocell. The photocell is directional and this direction is indicated by the green line. As the scanner sweeps across the light source, her circuitry is stimulated by the increasing light level to increase the speed of the drive motors, and simultaneously reduce the scanning speed. This behaviour gradually brings her closer to the light where her source of power resides.

On encountering an obstacle, the movement of her outer shell pressing against it, activates a 'trembler' switch. When this happens she enters an oscillatory state that switches rapidly between pushing and turning, "The steering-scanning motor is alternately on full- and half-power and the driving motor at the same time on half- and full-power" [9]. This enables her to wriggle free of the offending obstacle both in reality and in simulation, as shown in figure 2. She has separate drive and steering motors both ingeniously connected to the front *scanning* wheel. These run at different speeds according to her current state.

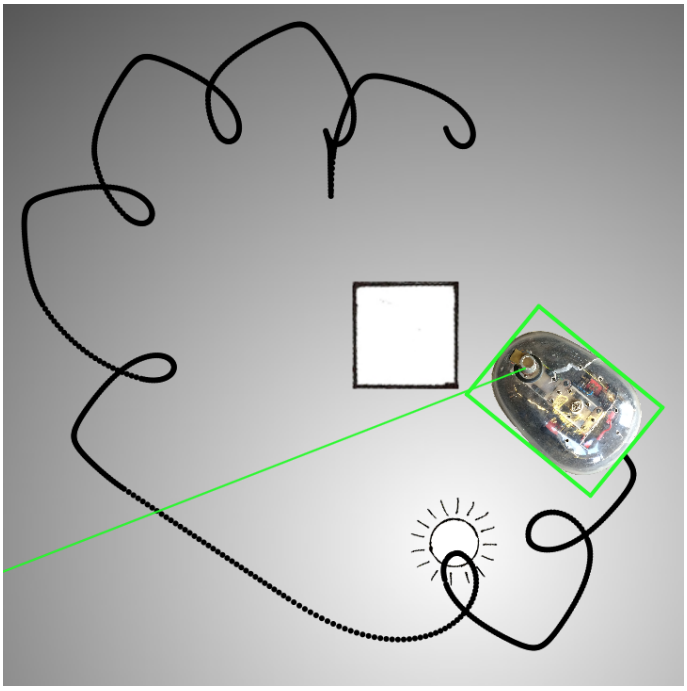

Figure 2. ELSIE simulation. Obstacle detection is simulated by the intersection of the robot's bounding box with that of an obstacle. It remains active for a short period after breaking contact.

Even where behaviour is continuous and analogue, it may be classified symbolically. This includes both sensor data and motor activity. The underlying analogue circuitry of ELSIE [9] is discretised into a set of symbols (E, P, N, O, R). These symbols are arranged in a string where the left-right order represents her observable behaviour in the world as time passes. We record only changes in state, so no pairs of neighbouring symbols are alike. To emphasise viable behaviours, the simulation includes a virtual battery that rapidly runs down as ELSIE explores (E) her environment. As she encounters obstacles (O), ELSIE wriggles free of them. ELSIE is attracted to light from afar (positive phototaxis P), but as she approaches on a full battery she is repelled from the bright light (negative phototaxis N). Only when her battery discharges is her repulsion to bright light overcome, allowing her to approach the light source and recharge (R). While ELSIE is recharging both motors are disconnected. These rules are summarised in table 1. The simulation is used to generate data for training and testing.

Table 1. ELSIE behavioural patterns

behaviour pattern	**collision**	**light level**	**scan/drive**
E - Exploration	no	low	fast/slow
P - Positive phototropism	no	medium	stop/fast
N - Negative phototropism	no	high	slow/fast
O - Obstacle avoidance	yes	ANY	1Hz oscillation
R - Recharge	no	ANY	stop/stop

From an enactive perspective, this is sensorimotor *data* only from an observer's perpective. An *enactive* system has no symbolic input/output. Of course a robot has sensors and actuators, but these can be thought of as acting, and being acted upon, causally. The photocel and trembler switch are simply non-symbolic components in an electrical circuit. There is no information passed from the external world to some notional information processor. Varela admits causal *perturbation* of an autonomous system, which we may think of as a pattern generator adapted to its environment.

A modern computer-based robot could collect sense-data. Within an enactive architecture it is important that this raw data is stripped of any intrinsic meaning. Meaning only arises when sense data is combined and compared, and can be modulated by motor activity. Computer based systems that receive pre-labeled input are enabled in their narrow task domain by getting this semantics for free, but are denied the opportunity of discovering a semantics of their own.

3 AUTONOMOUS SYSTEMS

While the behaviour of robots like ELSIE could be captured as a dynamical system, Varela sought alternative approaches that could be used to understand the cyclic nature of autonomous systems. In Principles of Biological Autonomy [14], Varela explores the logic of distinctions expounded by G. Spencer Brown [13], using this to discover autonomous recurrent states. State-machines provide a similar alternative that are perhaps more familiar to a modern audience. Representing a state-machine in matrix form as an adjacency matrix, enables us to analyse these cyclic behaviours, or *eigenbehaviours*. This eigenbehaviour is the signature of the autonomous system, constituting its very identity.

Each node of a state-machine represents a distinct, separate state, which is simple and effective for primitive robots, but with increased complexity, the method would quickly get out of hand because of combinatorial explosion. As an autonomous system a state-machine

satisfies the need for *organizational closure*. This means that there is no input/output to consider, so the transitions of the state-machine are unlabelled and simply connect one state to another to form a graph that closes in on itself. It is state-driven and finite, and is therefore capable of generating cyclic behaviour. For this analysis we are interested in the fundamental 'loopiness' of the behaviour it is capable of generating.

The nature and number of states is not to be identified with the behavioural states of the observable symbolic explanation above. The states are hidden, in that the states that the autonomous system passes through are not directly apparent in the symbolic behavioural data. State is hidden in the same sense that it is hidden in a Hidden Markov model, but for the analysis of the autonomous system used here we do not require the probabilistic parameters (transition and output probabilities) of the HMM.

A state-machine as a directed graph can be represented as a square adjacency matrix, where a one at the intersection of a particular row and column represents a possible transition from the current state (row) to the next state (column). The coupling between the autonomous system and its environment, or at least a symbolic description of it, is provided by another matrix that relates the hidden states of the autonomous system to the behavioural sensorimotor symbols. This is an incidence matrix relating hidden states to output symbols. The sensorimotor symbol can be described as a function of the current state, permitting multiple states to emit the same symbol but without the additional variability that HMMs allow for with a vector of output probabilities associated with each state.

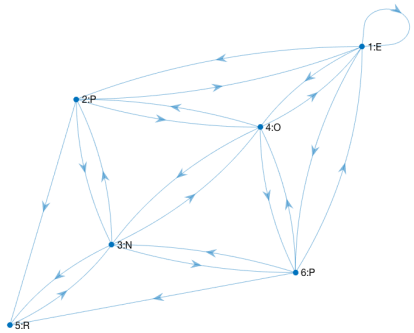

Figure 3. A 6-state machine produced by generating a Hidden Markov Model and discarding the transition probabilities has more edges than necessary. The graph shows each node number and emission symbol.

The adjacency matrix can be constructed using the tools for training Hidden Markov Models. Given one or more symbol sequences (emissions) it is possible to estimate the transition probabilities for a hidden Markov model using the Baum-Welch algorithm [2], a hill-climbing search that will converge to a local maximum from an initial randomised HMM (using the MATLAB function *hmmtrain*). The emission matrix can be initialised to a random incidence matrix with a functional mapping from states to emission symbols (a single one in each row). The transition probabilities it returns may be converted into *possibilities* by mapping all positive probabilities > 0, to 1. However, it seems wasteful to calculate the full transition matrix and then discard the probabilities in this way. There is a further issue with this approach. In addition to this wastage, graphs generated in this way have more edges than necessary to capture the behaviours in the training set, see figure 3 for an example.

An alternative *edge-removal* hill-climbing algorithm (see APPENDIX A) is able to generate the adjacency matrix directly. The intuition here is that we can begin with a matrix of ones, and then knock-out the edges (flipping matrix elements from 1 to 0) at random, subject to a check that this hasn't eliminated the potential for some observed behaviour. In fact, as the graph is not reflexive the main diagonal can be knocked out from the start. At each step the training set is consulted to ensure that all observed behaviours remain a possibility. If the adjacency matrix fails the test, then the change is reversed. This process continues until no more elements can be changed without failing the test. As before, the emission matrix is generated randomly and is held constant throughout. This produces the leaner graph of figure 4.

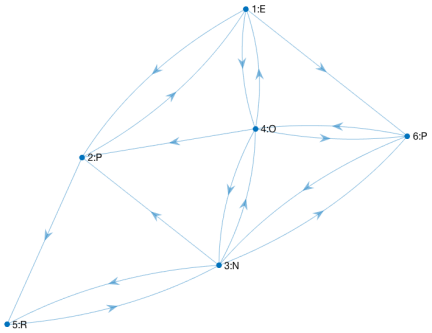

Figure 4. This 6-state machine generated by edge-removal hill-climbing captures complementary contexts of nodes emitting identical symbols, 'P'.

To test that a state-machine can potentially generate a behaviour, it must be treated as a non-deterministic finite-state-machine. At any given time it can be in multiple states; all the states that it may potentially have reached. In the initial configuration, any and all of the states are a possibility. As each behavioural symbol is consumed, some states are knocked out that cannot make that transition, and others are added as all valid transitions are followed. If at any point the state configuration is empty, then the state-machine fails the test.

For a given graph size, the edge-removal algorithm produces many candidate graphs each with a different random emission matrix as a starting point. How might we select among them? A measure is required that enables us to maximise the potential for cyclic behaviour. With an autonomous system captured in this form it is possible to carry out an analysis based on the number of walks that can be made from any node back to itself. From a range of graph centrality measures, *subgraph centrality* [7] emphasises the cyclic, oscillatory patterns of behaviour that are the hallmark of autonomy [1].

Subgraph centrality is based on the number of closed walks from

a node back to itself, a closed walk being a succession of edges starting and ending at the same node. 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. Subgraph centrality [7] is defined as the weighted sum of the closed walks of length k starting and finishing at a given node. The subgraph centrality, SC, of A for node i is defined in equation 1 below. 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 is equivalent to the diagonal entry of the matrix exponential of the adjacency matrix, e^A [8].

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

The Estrada index [4, 6] is defined in equation 2 below, to be the sum of the elements of the subgraph-centrality. This is equivalent to the *trace* (sum of the diagonal) of the adjacency matrix exponential.

$$EE = tr(e^A) \tag{2}$$

The *natural connectivity* of an autonomous system can be understood 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 the *natural connectivity* of a graph [11], defined in equation 3 below, where n is the number of nodes (hidden states).

$$\bar{\lambda} = ln\left(\frac{EE}{n}\right) = ln\left(\frac{tr(e^A)}{n}\right) \tag{3}$$

The natural connectivity of the graph has the nice feature that it changes monotonically with the removal (or addition) of edges[11, 16]. As the graphs produced by edge-removal have fewer edges than those derived from the Hidden Markov Model, they have correspondingly lower indices.

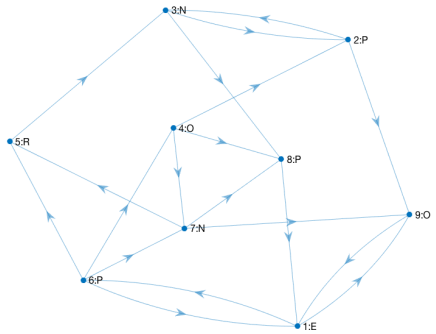

Figure 5. A 9-state machine generated by edge-removal with *minimal* natural connectivity (min-connectivity).

During the edge-removal search, the natural connectivity index falls monotonically with each edge removal. As this index is independent of the emission matrix and depends only on the adjacency matrix, it allows us to compare alternative solutions. It is worth reiterating that this gives us a measure of the robustness of an autonomous system, *without regard to its behaviour*. The focus is entirely on a system's ability to maintain its organisation seen as a recurring process. Non-connected graphs are eliminated at this stage as the Estrada index depends only on the Estrada indices of its connected nodes [5]. Of all the remaining minimal graphs, it allows us to discover the most robust solution with maximal natural connectivity. The 6-state solution of figure 4 was found within 1000 independent trials.

The interesting thing about the graphs of both figure 3 and figure 4 is that in both cases the search appears to have 'discovered' the underlying "EPNPE" cycle that occurs while the robot, ELSIE, is orbiting around the light source. As her turret rotates it transitions from a low light-level (E) while facing away from the light, through an intermediate spike of medium level light (P), to bright light (N), and then back again (P) as she veers away from the light. These state-sequences are primarily a feature of her electro-mechanical construction, rather than her electronic circuitry. As two of the hidden states map to 'P', more of the different contexts of these two 'P' emitting states is captured in the graph.

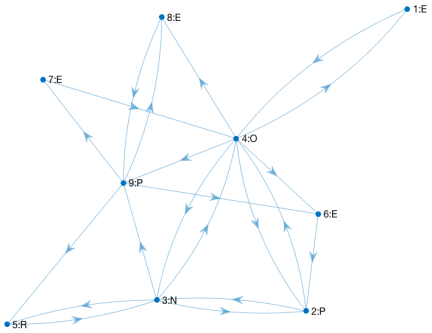

Figure 6. A 9-state machine generated by edge-removal with *maximal* natural connectivity (max-connectivity).

What does it mean in practice to maximise natural connectivity? The edge-removal algorithm destroys a huge amount of robustness and potential behaviour, so we're working at the fringes with resource limited systems having a fixed set of states and only a minimal number of state-transitions. What is the virtue of maximising natural connectivity – retaining whatever potential the system has – rather than further minimisation of natural connectivity at this stage? An example of a graph with minimal natural connectivity ($\bar{\lambda} = 0.3836$) is shown in figure 5. Call this the minimal graph. This is a graph with 9 nodes mapping onto the same set of 5 behavioural patterns, so we see more nodes mapped to the same output. Contrast this with the 9-node graph with maximal natural connectivity ($\bar{\lambda} = 0.9059$) in figure 6. There's no discernible difference between the in or out-degree

of either graph. The minimal graph has mean in and out-degrees of 2.11 (they are the same), while the maximal graph has mean in and out-degrees of 2.33.

The clearest difference between these graphs is in the number of short-cycles evident in their structure. This should come as no surprise as this is precisely what the Estrada index measures, and indeed it gives more weight to these shorter cycles. As we saw above, 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. The trace of A^2 divided by 2 is the number of directed digons (2-sided directed polygons in the graph), while the trace of A^3 divided by 3, is the number of directed triangles. The division is necessary because we can walk around the path starting at any of its vertices. The $tr(A^2)/2 = 3$ digons in figure 5, and $tr(A^2)/2 = 6$ digons in figure 6, may easily be counted. The pattern doesn't straightforwardly extend to squares as digons also produce 4-cycles. However, the Estrada index is not concerned with whether such 4-cycles are generated by squares or combinations of digons. Table 2 informally summarises short 2, 3, 4, 5 & 6-cycles found in these examples of minimally and maximally connected directed graphs. It can be seen that the tendency in maximally connected graphs over minimally connected graphs is towards these cycles. Both graphs are sufficient to produce the observed behaviours in the training set, so it is plausible that the reduced performance of the minimal graphs is due to over-fitting.

Table 2. n-cycles in minimally and maximally connected directed graphs.

graph	$tr(A^2)$	$tr(A^3)$	$tr(A^4)$	$tr(A^5)$	$tr(A^6)$
min-connectivity	6	0	22	20	72
max-connectivity	12	15	64	140	435

If we were to consider robot phenomenology, then this autonomous system defines the world as seen from the robot's perspective. It defines how the robot *enacts* the world. Of course, not any behaviour goes; the autonomous system must be adapted to its environment in a way that produces effective, or viable behaviour (survivability over some time-scale). However it does not simply embody a model of its environment, but brings meaning to its environment grounded in its physical, existential needs.

4 EXPERIMENTAL RESULTS

To validate the autonomous system against the symbolic data of recorded test behaviours, we may evaluate the state-machines produced by the edge-removal algorithm against the test-set of behaviours. As in the algorithm itself, we test for the *possibility* that all sequences could be produced by the state-machine. This generates a binomial distribution of machines that either pass or fail the tests. we can plot the data in a bar-chart assigning the data to one of ten bins based on the natural connectivity score. If we count a pass as +1 and a fail as -1, and sum all the entries in each column we will see the bar rise above zero if there are more passes on balance, or it will drop below the zero line if there are more failures. If natural connectivity has no bearing on this then we would expect the passes and failures to be evenly distributed. However, the distribution is far from even as can be seen in figure 7 a representative example based on an 8-state machine. It looks like the machines with minimal natural connectivity are over-fitted to the data, while there's a window for those machines at the top end with maximal natural connectivity – greater redundancy – to pass; accounting for both the training data and the test data.

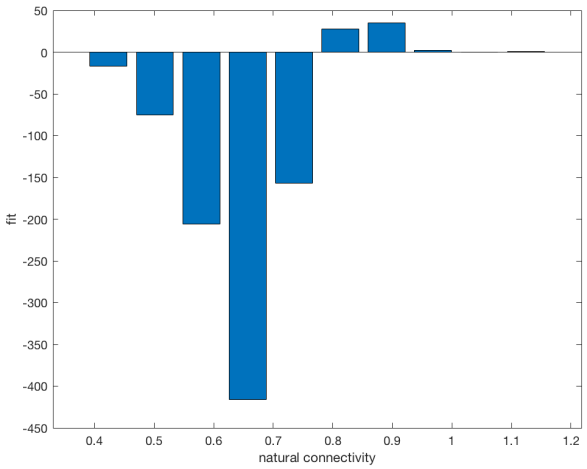

Figure 7. Bar-chart of the natural connectivity of an 8-state machine, plotted against the summed binomial data (± 1) representing a pass/failure to account for the test data. This shows an uneven distribution that tips from overall fail to overall pass around a natural connectivity of 0.81.

This pattern holds for all the graphs explored between 6 and 10 nodes, although the turning point between overall pass/fail varies. These values present as the curve in figure 8 asymptotically approaching a value around a natural connectivity of 0.78.

The hypothesis is that there is a correlation between the natural connectivity of an autonomous system, and the corresponding success rate when evaluated against the test set of behaviours. It isn't immediately obvious that this should be the case, as natural connectivity is a function only of the graph topology. We are really asking if the qualitative features of the graph favoured by natural connectivity lend themselves to good discrimination when it comes to testing.

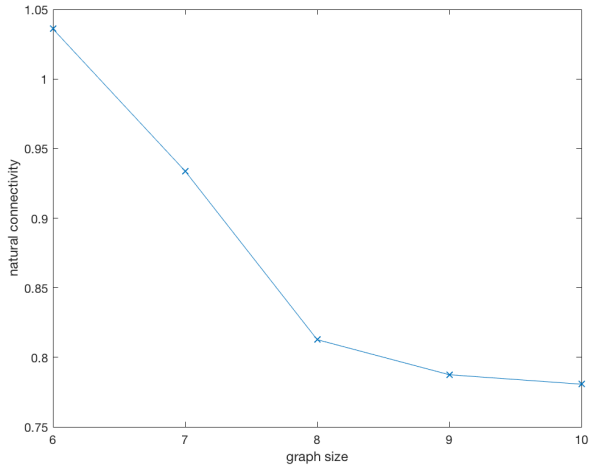

Figure 8. Curve of the tipping point from overall fail to overall pass as we explore machines with increasing number of states.

For each of the graph sizes between 6 and 10 nodes, the data was subdivided into two bins at the threshold defined by the curve in figure 8. We want to confirm that the binomial distributions are different; those below the threshold being predominantly failures, and those above, predominantly passes.

For each of the given graph sizes, 10K independent trials of the edge-removal algorithm are used to produce a state-machine, many of which are eliminated because they do not have a connected graph. The remaining machines are tested and assigned to the appropriate bin depending on whether they passed or failed. A Chi-squared test is applied to evaluate how likely it is that the observed differences between these two populations arose by chance.

The Null Hypothesis, $H0$, is that there is no difference between the two populations. For a range of networks explored, from 6 nodes to 10 nodes, the p-value is very close to zero, far less than a significance level of 5% (0.05). We therefore reject the Null Hypothesis and conclude that there *is* a significant difference between graphs with high natural connectivity (loopiness) and those with lower natural connectivity. Graphs with high natural connectivity appear to avoid over-fitting due to their built-in redundancy.

5 CONCLUSION

This paper explores the use of state-machines to capture operational models of autonomous systems. It explores the relationship between a sequential symbolic representation of a robot's behaviour, and an operational representation – a state-machine – that may be used to produce such a behaviour. We have explored a centrality measure on the graphs underlying such a state-machine model and found that natural connectivity measures the kind of *loopy* behaviour we would associate with an *autonomous* system. An edge-removal hill-climbing algorithm in conjunction with a natural connectivity measure of graph structure allows us to derive an adjacency matrix representation of the state-machine from observed data, minimising the chances of over-fitting to the training data. The surprising thing is that this final selection of candidates is made purely on the basis of the graph topology without regard to the training data; in other words, purely on the basis of its *operational closure* rather than input or output.

6 APPENDIX A: edge-removal algorithm

```
edge_removal(trials)
  score = 0

  for i = 1:trials
    a = ones matrix - leading diagonal
    e = random emission matrix (single 1 on each row)
    x = indices of ones in a
    while x is not empty
        i = index drawn from x
        a[i] = 0
        if all observations can be generated by a,e
            x = indices of ones in a
        else
            undo change to a
            x = x - i

    % natural connectivity
    n = log(trace(expm(adj))/m)
    if n > score and a is connected graph
        adjacency = a
        emission = e
        score = n

  return adjacency, emission, score
```

REFERENCES

[1] Steve Battle, 'Principles of robot autonomy', in *Philosophy after AI: mind, language and action, Proceedings of AISB 2018*, (April 2018).

[2] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, 'A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains', *Annals of Mathematical Statistics*, **41**(1), 164–171, (1970).

[3] R.A. Brooks, 'Intelligence without representation', *Artificial Intelligence*, **47**, 139–159, (1991).

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

[5] Zhibin Du, 'An edge grafting theorem on the estrada index of graphs and its applications', *Discrete Applied Mathematics*, **161**(1), 134 – 139, (2013).

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

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

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

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

[10] M. Heidegger, *Being and Time*, Library of philosophy and theology, SCM Press, 1962.

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

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

[13] G. Spencer-Brown, *Laws of Form*, Dutton, 1979.

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

[15] F.J. Varela, E. Thompson, and E. Rosch, *The Embodied Mind: Cognitive Science and Human Experience*, MIT Press, 1991.

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

AI-Generated Music: Creativity and Autonomy

Caterina Moruzzi[1]

Abstract. In this paper I discuss the impact of AI on one of the key topics in the philosophy of art: the nature of musical works. The question I will address is the following: 'Can a computer create a musical work?'. Due to its complexity and subject-dependency, the concept of creativity eludes an objective definition and quantification. With the aim of finding a non-arbitrary way of measuring creativity, in section 2 I individuate autonomy as a necessary feature exhibited by creative processes. In section 3, I investigate the case study of the generation of music by a new model of generative algorithms: Generative Adversarial Networks (GANs). I claim that the use of GANs in music composition may grant the software a sufficient level of autonomy for deeming it able to create musical works. In addressing the inherent difference of GANs from other kinds of software used in algorithmic composition, I will borrow some insights from the discussion on Integrated Information Theory. In section 4, I discuss the quantity of integrated information, or Φ ('phi'), that a system for music generation which makes use of GANs may possess and what this tells us about its levels of creativity[2].

1 INTRODUCTION

The last decades witnessed an exponential proliferation of AI music composition programs. The hard-coded algorithmic composition systems of the outset are progressively giving way to a more advanced use of neural networks and Deep Learning software, with a consequent increase of the sophistication and quality of the music produced. The Mozart and Lady Gaga of the future are a set of silicon chips: Jukedeck, Flow Machines, Aiva and other programs are gaining more and more followers in many music platforms, raising a growing enthusiasm and consent among their fans. While the latest developments in the field of AI-generated music are matter of excitement among both listeners and researchers, they also raise less practical and more philosophically oriented questions. In this paper I discuss the impact of AI on one of the key topics in the philosophy of art: the nature of musical works. The question I will address is the following: 'Can a computer create a musical work?'. The answer that we give to this question is interesting not only from a philosophical and theoretical point of view. Indeed, it carries also relevant consequences for considerations regarding copyright and intellectual property and, in addition, it can give us some insight into the nature of human creativity itself[3].

Before I can move on to the consideration of whether an algorithm can be deemed creative, I need first to define what creativity is. This is what I aim to do in the first part of the paper. In the literature, many definitions of 'creativity' have been given but none of them emerges as the paradigmatic one that can be employed in all fields and disciplines. My scope in this paper is not to overturn existing approaches to creativity, but instead to suggest possible avenues of exploration that may be beneficial for an understanding of human creativity. This will hopefully help the development of better software and of a smoother and more effective interaction between humans and machines. I will define creativity as a process which is necessarily autonomous. As I will explain in more detail, this is a minimal definition but it serves the scope of trying to find an objective measure for creativity in order to move on to the study of its many other characteristics.

In section 3, I investigate the role that the notion of autonomy plays in software for music composition. In particular, I examine a generative model for music generation: Generative Adversarial Networks (henceforth GANs) [23]. The process of creation undertaken by GANs displays a level of autonomy from pre-existent sets of data which is noticeably bigger in respect to other generative algorithms. Hence, I argue that GANs are a good candidate for being deemed creative. I focus on software for music-generation, since music involves a wide range of bodily, contextual, and technological aspects in which creativity is involved. At the same time, it is a particularly challenging field for algorithmic composition due to its temporal dimension and its multi-layered complexity. The focus on music generation does not prevent us from extending the results of this research to the analysis of other fields of application of creativity, though.

In the last part of the paper, I tentatively suggest a way to objectively measure the amount of creativity possessed by generative models, borrowing some insights from the discussion on the measurement of consciousness proposed by the Integrated Information Theory (IIT) [50]. This theory argues that the more a system is integrated, the more consciousness it possesses. I will discuss the quantity of integrated information, or Φ ('phi'), that a generative model for music composition may display and what this tells us about its creativity.

The period we are living provides us with exciting material and opportunities to research on the topic of creativity. I am confident that gaining insight into artificial creativity can yield some other, more encompassing, results regarding how the human mind works and how we can enhance its performance through the collaboration with technology.

2 CREATIVITY AS AUTONOMY

The topic of creativity raised the interest of many scholars in different fields. However, discussions on creativity usually end leaving participants even more confused on what the nature of creativity is. This is partly due to the fact that creativity is defined differently by people working in different fields.

Artists normally deem creativity as a property of 'genius' that only some people seem to possess, at least at its more distinguished level [1]. Computer scientists, for their part, consider creativity as something that can be recreated through an

[1] Dept. of Philosophy and Music, University of Nottingham, NG7 2RD, UK. Email: caterina.moruzzi1@nottingham.ac.uk.

[2] This research is funded by the Midlands3Cities/University of Nottingham.

[3] See section 5 for a discussion on this.

algorithm and they usually place more relevance on the final product rather than on the process that leads to it [8, 13]. Cognitive scientists look for neural correlates of creativity in the brain and try to understand the mechanisms underlying it [22, 47]. Philosophers conduct a meta-study of what researchers in other fields say about creativity. As a consequence, they describe creativity in a variety of ways: as a property of a process, a property of a product, an emergent entity, etc [40].

From this quick overview, it is clear that there is no consensus on what creativity is. The first step we need to take in the investigation on the nature of creativity is thus trying to shed light on the issue. Machine learning is achieving impressive results also in fields which were before considered as exclusively human, such as creative arts, but some people question the possibility of mechanising a human process that we do not fully understand [33]. The field of neuroscience is in constant development but we do not have a clear explanation of what is going on in the human mind when we undertake a creative activity yet.

In this paper I do not have the presumption of providing a definite answer on what creativity is, nor to indicate its neural correlates or the way in which to replicate it in an artificial substratum. Rather, my aim is to pave the way for one possible interpretation of creativity which may hopefully lead to interesting results if pursued further.

We arguably share the intuition that creative processes present something special in respect to other, more mundane, activities. This special feature of creativity has been interpreted in different ways, as value, originality, intentionality, etc. [3, 5]. Finding out which are the essential characteristics that make up the nature of creativity is the end result that I want to obtain. Thus, I am not in the position of stipulating at the beginning what these 'special features' are. Instead, I provide a minimal account of the creative process which is intended as a baseline, a way to determine the primary nature of creativity in order to proceed in the investigation and refine our picture of creativity as we gain more insights on it.

I suggest that creativity is the property of a process which is necessarily autonomous. The choice of focusing on the creative process instead of the product is motivated by the need to try and achieve a view on creativity which is as objective as possible. Addressing out attention to the product that originates from a creative process, does not provide us with any insight regarding how creative thought originates or what it entails [54]. The consideration of a product as creative is not objectively measurable but, rather, it depends on how we perceive it and on a series of contextual factors [14, 18][4].

I believe that the exploration of the mechanisms underlying a creative process may instead be more liable to an objective investigation. As I understand it, the creative process is not necessarily individualistic. The Romantic notion of creativity that permeates our conception of it obscures a more communal idea of creativity. Creativity is a process that can be performed with the collaboration of and influence from other people and contextual elements. Most importantly, creativity is not exclusive of an elite of 'geniuses' but it is common to every human, in various degrees[5].

I defined the process of creativity as 'necessarily autonomous'. Autonomy is widely recognised as an essential property of creativity [9, 15, 30, 31]. However, autonomy can be variously defined as the generation of inner goals [30], the ability to respond to known and unknown inputs [7], or the freedom to generate ideas [16]. A possible concern is that autonomy is too strong a requirement for a process to be creative. Also humans, in fact, are not completely autonomous but instead constrained by their body, social context, the tools they use, etc. In this paper, I understand autonomy not as complete freedom but as not being completely reliant on a given set of data [31].

In the case at issue, thus, creativity is understood as an autonomous process of creation of an output which is not limited to the imitation of a given corpus of music. I pre-empt a possible objection here: also human musicians rely, to a certain extent, on a set of data and instructions. I grant this is the case, however, once the learning stage is complete, the process of creation may lead in directions which depart from the training set. This is the kind of autonomy I refer to when analysing the possibility for a software to be creative.

A last specification regarding autonomy is needed. It is possible to have agency without displaying autonomy [26]. As defined by the Oxford English Dictionary, agency is an 'action or intervention producing a particular effect'. It is thus possible for an algorithm to have agency, namely to have an effect, without being autonomous. This distinction is essential for determining the difference between computer-assisted and computer-generated music. In computer-assisted music, the software does not necessarily need to be autonomous, but instead it needs to have agency and provide a contribution to the final outcome [24, 25][6]. Despite the fascination of the field of human-AI collaboration in music production, in this paper I am interested in the more radical case of computer-generated music. In this latter instance, computers do need to display autonomy, and not only agency, in order to be deemed creative.

In the Introduction I anticipated that I would have focused on a particular instance of generative software: GANs.[7] The reason for exploring this model instead of others is that I believe their process of creation is inherently different from other software for music generation. As I will explain, the interplay between generator and discriminator in the structure of GANs grants the process of creation a considerable level of autonomy from the training set. GANs have already reached noticeably better results than other algorithms in the field of visual arts.[8] The application to music is more recent and, due to the complexity of musical structures I mentioned earlier, the outcomes obtained are less

[4] The subject-dependence of the notion of creativity as applied to a product is also the reason why the Turing Test is deemed by many a non-reliable measure of creativity [57, 9, 3].

[5] I will not enter the debate of whether also animals are capable of creativity here. For a discussion on this theme, see [27].

[6] See also [55] for a discussion on joint-authorship and the necessity to define creativity to determine the ownership of copyright in the case of human-AI interaction in the generation of music.

[7] For an extensive overview on algorithms for music generation, see [38]. For GANs, see [23].

[8] See https://www.theguardian.com/technology/2019/mar/04/can-machines-be-more-creative-than-humans and https://www.theatlantic.com/technology/archive/2019/03/ai-created-art-invades-chelsea-gallery-scene/584134/ for recent examples of the success of GANs in the visual arts.

sophisticated. However, I believe that, with future developments, GANs will be able to obtain impressive results in music, comparable to the ones they already achieved in visual arts.

3 GANs AND CREATIVITY

Generative models are algorithms developed with the aim of analysing and understanding data from a pre-existent set. Given a label or a hidden representation, they can predict the associated features and generate new data similar to that provided by the training set. Generative algorithms have been mainly used for the classification and generation of images but recently they started to be applied also to the generation of music [16, 19, 58].

Generative Adversarial Networks (GANs) have been introduced in 2014 as a new kind of unsupervised generative algorithm [23]. GANs are composed of two neural networks: a generator and a discriminator. The generator has the role of originating new data instances, while the discriminator evaluates them for authenticity. This model is defined as 'adversarial' because generator and discriminator are pitted one against the other in what Goodfellow describes as a game of cat and mouse: 'The generative model can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, while the discriminative model is analogous to the police, trying to detect the counterfeit currency. Competition in this game drives both teams to improve their methods until the counterfeits are indistinguishable from the genuine articles.' [23: 1]

The aim of the generator is to fool the discriminator into thinking that what has been produced is a sample which is part of the training set. The aim of the discriminator is to catch the generator out any time it produces a fake sample. The two neural networks, generator and discriminator, are trained simultaneously in a minimax two-player game, in a competition to improve themselves. The ideal solution is that the discriminator always outputs the chance 0.5 that the input coming from generator is real. This would mean that the generator has learnt to produce output indistinguishable from the training samples and that the discriminator is not fooled into thinking that it is false but leaves instead open the possibility it can be either authentic or fake.

The main focus of GANs is to generate new data from scratch which is indistinguishable from the data in the training set. The generator alone would create just random noise. The role of the discriminator is to guide the generator, providing feedback to create data instances that look (or sound) like the training samples. The feedback loop occurring between generator, discriminator, and data set is what allows both the generator and the discriminator to improve their performance[9].

The mechanism applied by GANs for generating new images and new music is fundamentally the same. In the case of music random noise is used as the input to the generator, whose output are melodies [9: Chapter 5][10]. Another model which is frequently used for music generation is Recurrent Neural Networks (RNNs). RNNs are neural networks which learn a series of items thanks to recurrent connections. In this way 'the RNN can learn, not only based on the current item but also on its previous own state, and thus, recursively, on the whole of the previous sequence. Therefore, an RNN can learn sequences, notably temporal sequences, as in the case of musical content.' [9: 65].

The capacity of RNNs to learn sequences has made them one of the preferred models to generate a temporal output such as music. As I will discuss in section 3.3, RNNs produce arguably better results than GANs [11][11]. However, their structure is inherently different. The inner feedback loop present in GANs provides them with an autonomy which is not available to other models. The fitness function (a score to evaluate how close the output comes to meeting the specification of the desired solution) in GANs can indeed come from within the system, not from external feedback [10]. The interplay between generator and discriminator is what distinguishes GANs from other algorithmic models and, from it, two major benefits emerge: (1) the capacity of GANs of self-evaluation, and (2) their relative autonomy from a pre-existing set of data. I will examine each of these benefits in turn.

3.1 Self-evaluation in GANs

Self-evaluation, namely the capacity of making normative judgements regarding the output we produce, has been indicated by some as an essential feature for creativity [21]. When engaging in a creative process, we normally reflect on what we produce and try to improve according not only to the feedback that we receive from the outside but also according to the inner feedback we provide ourselves with. Self-evaluation is a process that calls into questions many other activities and properties of our mind, like consciousness and intentionality. It does not surprise, then, that we intuitively struggle to ascribe self-evaluation to artificial machines: 'Although many generative models have become quite sophisticated, they do not contain an element of reflection or evaluation, and therefore might not necessarily be considered creative systems in their own right' [2: 2]. However, if we take away other, notoriously complex, notions such as consciousness and intentionality from the definition of self-evaluation, we can understand it as the capacity to improve the quality of the outputs based on the consideration of past performance.

The feedback loop that represents a vital element of the architecture of GANs, can be interpreted as a self-evaluation process. The interplay between the generator and the discriminator, indeed, maps the human creative process of trial and error. What is especially relevant, is that this feedback mechanism happens within the GAN and does not involve external players. Interactive evolutionary computation traditionally makes use of human judges to gain evaluation in the cases in which a fitness function is not known or hard to determine (like in the case of aesthetic qualities in the arts) [21, 45]. In these cases, the feedback here comes from the outside, so it cannot be classified as self-evaluation. In GANs, instead, the feedback comes from within. This, I argue, is a benefit that GANs have in respect to other algorithmic models, since it contributes to the overall autonomy of their process of creation and, hence, to their creativity.

A potential objection may be raised here: GANs are not really autonomous because they rely on a pre-existent training set. I

[9] In this paper I do not provide the technical details of how GANs work. For more details see [10, 23].

[10] I will come back to the discussion on the quality of music generated by GANs in section 3.3.

[11] However, RNNs are not exempt from problems, see [39].

believe this argument is not against the autonomy that can be attributed to GANs. Indeed, also humans rely on a set of training data and instructions for every creative action they undertake [53]. What is relevant in order to attribute a certain level of autonomy to the system, is that it displays independence after the learning stage has been completed and during the process of creation.

Lastly, there is another attribute of GANs that can vouch in favour of their creativity. The kinds of learning on which machine learning has focused so far are: rote learning, learning from instructions, learning by analogy, learning from examples, and learning from observation [46]. The most common kind of learning applied in generative models is learning from examples. I argue that the generator in GANs does more than simply learning from examples, it learns *by doing* [4]. The learning process of the generator in fact relies heavily on the competition it plays with the discriminator. It improves its performance by pitting against the discriminator and producing an output that is every time better than the last. This is similar to the process of learning also humans go through when engaging with a new activity or topic. Sure, algorithms still lack human-like embodiment, an element which plays an essential role in the human process of learning [29, 48, 54]. Yet, with the advancements in the field of robotics it is not excluded that they may achieve perceptual features and skills in the future.

Self-evaluation and learning by doing have been individuated as two features of GANs which advocate in favour of their possibility of engaging in creative processes. In what follows I discuss a further benefit of GANs: their relative autonomy in respect to a pre-existent set of data.

3.2 Autonomy from pre-existent corpus

The majority of software for music composition is programmed to imitate the style of the music provided in the training corpus[12]. Imitation may not be deemed enough in order to recognise autonomous creativity to a system, though[13]. A recent evolution of GANs is trying to achieve the necessary detachment from the training set in order for its output to be not a mere imitation of the samples provided during the training stage, but instead a more autonomous creation. ADD This model, called Creative Adversarial Networks (CANs), aims to deviate from the training set and to create a new style: 'The network is designed to generate art that does not follow established art movements or styles, but instead tries to generate art that maximally confuses human viewers as to which style it belongs to.' [17: 5] The authors describe the mechanism of CANs by saying that it 'tries to increase the *stylistic ambiguity* and deviations from style norms, while at the same time, avoiding moving too far away from what is accepted as art. The agent tries to explore the creative space by deviating from the established style norms and thereby generates new art.' [17: 5]

An interesting result of CAN is that, without human intervention, the algorithm deviated from the style norms of the training set to generate abstract paintings. This has been interpreted by the creators of CANs as a parallel between the trajectory of human art history and the path undertaken by CANs: in both cases, the agents 'opted for' abstraction.

How to define what a style is, is a challenge in itself [32]. For the sake of this discussion, I understand a style as a recognisable pattern which is replicated in subsequent instances and which has elements of uniqueness. This definition is undoubtedly vague and open to criticism[14]. My aim here is not to determine what can be identified as a style, though. Instead, I wish to acknowledge the fact that the production of output which displays certain uniform characteristics which make it distinguishable from other instances (what I define a 'style') can be interpreted as a measure of creativity of a system.

The creation of a new recognised style has not been achieved by CANs, yet. However, I argue that CANs are on the right track to develop a system which is increasingly autonomous from pre-existent corpus of data and from external feedback and, thus, which can arguably display the essential features for a system to be deemed creative.

3.3 Limitations of GANs

GANs brought about a revolution in generative algorithms, allowing to achieve results that were unthinkable before. However, they are not exempt from criticism, mainly for two reasons: they are very hard to train and (in the generation of music) they produce worse results than other algorithmic models.

Training GANs requires finding a Nash equilibrium between generator and discriminator[15]. However, GANs are usually 'trained using gradient descent techniques that are designed to find a low value of a cost function, rather than to find the Nash equilibrium of a game.' [42: 1] As a consequence, training is often unstable[16]. The interplay between generator and discriminator, the feature that constitutes an advantage for GANs and their creativity, also causes its difficulties in the training stage. Various solutions have been proposed to face the difficulty of training GANs but none of them seems to be particularly successful[17].

A second difficulty that GANs face in their application to the generation of music is the arguably poorer quality of their output in respect to other algorithms for music generation. For example, RNNs, the other model for music generation I mentioned, are able to produce music which displays a more consistent structure and a melody which is easier to follow and better balanced than GANs[18]. Nevertheless, I argue that this does not diminish the

[12] For example, FlowComposer, DeepBach, EMI, COMPOSER, and GenJam are based on an imitation principle.

[13] Even though, it may be argued, also many human composers, at least to a certain extent, compose by imitating the music created by others.

[14] A concern related to the definition of 'style' is the fact that, as the notion of creativity, also the notion of style can be partly subject-dependent. Moreover, a question that can be asked is whether a style needs necessarily to be 'intentional', namely if an agent needs to have the intention of creating a style or whether it can happen by chance.

[15] In game theory a Nash Equilibrium happens when one player will not change its action regardless of what the opponent might do. In this case the two players are generator and discriminator.

[16] Other difficulties in training GANs include the problem of vanishing gradient and mode collapse. For technical details, see [10, 39, 42], and https://medium.com/@jonathan_hui/gan-why-it-is-so-hard-to-train-generative-advisory-networks-819a86b3750b.

[17] They include techniques such as minibatch discrimination [42], DCGANs [41], and Wasserstein distance [16].

[18] Examples of music composed with RNNs can be found at: https://magenta.tensorflow.org/performance-rnn, http://www.hexahedria.com/2015/08/03/composing-music-with-

level of creativity that we can recognise to GANs. In fact, as I stated at the beginning, creativity needs to be measured on the basis of the process, not on the basis of the quality of the product.

Human evaluation has been extensively employed to judge the level of creativity or the quality of the output of genetic algorithms for music generation [16, 17, 28, 42]. However, having a human feedback on the creativity of the system might be useful to improve the system itself but, I argue, it is not useful to investigate the notion of creativity. The interpretation of the quality and creativity of an output is subjective and depends on a series of personal and contextual factors [57]. Moreover, the quality of an output is not dependent on the creativity of the process that led to its creation. If an art student produces a drawing whose quality is worse than the quality of a drawing by a professional artist we would not question the creativity of the art students but, rather, her lack of experience or technique. Similarly, we cannot judge the creativity of an algorithm on the basis of the quality of its output.

In the next section, I tentatively suggest a way to measure the creativity displayed by GANs. In order to do so, I will use the tools provided by the Integrated Information Theory of consciousness.

4 INTEGRATED INFORMATION THEORY TO MEASURE CREATIVITY

Integrated Information Theory (IIT) was proposed by Giulio Tononi in 2004 as a theory to explain and measure consciousness [49] It is described as 'an evolving formal and quantitative framework that provides a principled account for what it takes for consciousness to arise, offers a parsimonious explanation for the empirical evidence, makes testable predictions and permits inferences and extrapolations.' [52: 5]

In this section I suggest that we can use the tools offered by IIT to explore the possibility of finding an objective measure to creativity. I argue that this is possible given the link that occurs between consciousness, intentionality, and creativity. I start with the assumption that intentionality and consciousness are interconnected [3, 37]. Creativity is linked to consciousness as it is recognised as involving both conscious and unconscious processes of the mind [56]. Moreover, intentionality has been individuated as an essential feature of creative activities by many [3]. An agent necessarily needs to intentionally engage in an activity for this activity to be considered creative. Given the connection existing between these three concepts, I believe that a study on the nature of consciousness can yield some results also in relation to the nature of creativity.

One of the difficulties that I highlighted at the beginning regarding the study of creativity is its subject-dependent nature which eludes an objective measure. The possibility of quantifying consciousness offered by IIT may provide us with the instruments to bypass this obstacle and acquire a neater understanding of the nature of creativity.

IIT proposes a set of axioms and postulates to define what consciousness is. Consciousness is described as existent, structured, specific, unified, and definite [52]. The two attributes of consciousness that mostly interest us in the context of this discussion are 'specific' and 'unified'. 'Specific' refers to the cause-effect structure displayed by a system. This represents the 'information' part of IIT: 'information refers to how a system of mechanisms in a state, through its cause–effect power, specifies a form ('informs' a conceptual structure) in the space of possibilities.' [52: 8]

'Structure' refers instead to the intrinsic irreducibility of the system, which can be measured through the mathematical quantity Φ (phi): 'a conceptual structure completely specifies both the quantity and the quality of experience: how much the system exists — the quantity or level of consciousness — is measured by its Φ_{max} value — the intrinsic irreducibility of the conceptual structure; which way it exists — the quality or content of consciousness — is specified by the shape of the conceptual structure.' [52: 9]

The more the parts of a system are interconnected, the higher the phi. A high level of information means that the cause-effect powers of a system specify 'which out of a large repertoire of potential states could have caused its current state.' [51: 300] High integration means instead that 'the information generated by the system as a whole is much higher than the information generated by its parts taken independently. In other words, integrated information reflects how much information a system's mechanisms generate above and beyond its parts.' [51: 300] Consciousness is a fundamental quantity, just as mass, charge, or energy [50: 233]. The consequence presented by IIT is that every system which is minimally specific and unified, possesses a degree of phi and, hence, is conscious [50: 236] [19].

I argue that the inner structure of GANs is a good candidate for presenting an appreciable amount of phi. As mentioned, phi is measured according to how much a system can have an intrinsic cause-effect architecture and how much integrated information it presents. The interplay between generator and discriminator in GANs, and the feedback loop that allows them to improve their performance, constitutes a cause-effect structure which has powers both within and outside of the system. The integrated information of GANs derives from the causal interaction between the components of the system and from their level of irreducibility[20].

Consciousness in IIT is attributed to re-entrant systems. On the other hand, a feed-forward system that performs in the same way as a conscious human would only simulate consciousness, not realise it [20]. This is consistent with the idea expressed throughout this paper that creativity can be identified in the process rather than the product. No matter how similar a musical piece composed by an algorithm is to a piece composed by a human, if the process followed by the algorithm to generate it is not integrated and autonomous, it cannot be deemed creative.

recurrent-neural-networks/ and http://people.idsia.ch/~juergen/blues/. Examples of music composed with GANs can be found at: http://mogren.one/publications/2016/c-rnn-gan/ and https://salu133445.github.io/musegan/results.

[19] For other, more technical details on IIT, see [49, 50, 51, 52].

[20] IIT as applied to GANs can also be used to judge whether humans, whose causal power is represented by the training set and possibly by external feedback, are a necessary components of the system. Φ_{max} calculates to what extent the cause-effect structure of a system changes if the system is partitioned. The level of phi can, thus, be calculated with and without the human component to check which is the maximally irreducible section of the system with the highest level of phi.

What I presented in this section is a tentative suggestion on how to use tools provided by a neuroscientific theory to achieve a better understanding of creativity. However, this research is not free from difficulties. Measuring phi is very hard, as immense computational power is required to calculate all the relations within the cause-effect structure of a network. Despite this difficulty, calculating phi may turn out to be an achievable task as the computational power of computers is increasing exponentially over time. Either way, I am confident that the application of the theoretical framework offered by IIT may prove to be extremely beneficial to the study of a central feature of the human mind, creativity, and of its potential replication on artificial substrata.

5 CONCLUSIONS & FUTURE WORK

In this paper I discussed the question of whether algorithms for music generation can be deemed creative. I analysed GANs as a case study and I concluded that, thanks to their relative autonomy from a pre-existent set of data and to their capacity of self-evaluating their performance through a feedback loop between generator and discriminator, they are a good candidate for being considered, at least minimally, creative. In the last section of the paper I tentatively suggested to use the tools provided by the Integrated Information Theory of consciousness to measure the creativity that can be displayed by a system. Also in this case, given the causal power exercised by the inner components of GANs and their interconnection, I proposed that they may exhibit a considerable level of phi, and thus of creativity.

In this concluding section, I wish to indicate a domain in which the results obtained from it can play a significant role and possible developments of this research.

First, it should be noted that the investigation on creativity is not limited to the field of the arts. Creative processes are common to many other disciplines, from technology, to scientific discovery, to social creativity [48]. Achieving a better understanding on what creativity is can be especially beneficial for discussions in relation to copyright and intellectual property. Legal considerations on authorship need to address the question of what is required in order to get copyright authorship. Together with originality and novelty, creativity is a feature that is deemed necessary by law in order to recognise a product as worthy of copyright protection [8]. The problem is that creativity has not been fully defined by copyright law [8, 55]. This vagueness leads to significant uncertainty especially when addressing the issue of creativity in AI. In the last years there have been many cases of human-AI collaboration in the generation of music. What needs to be established, then, is the role played by those contributing to the final product: human musicians, programmers, and the software itself. An informed discussion on creativity in AI and on the autonomy of AI agents in respect to humans may help settle the issue of joint authorship and copyright [55].

I conclude by pointing out a further line of research that may emerge from the discussion conducted in this paper regarding creativity and AI. I start with a provocation: even if AI could reproduce the human creative process in music and generate a creative music product, exactly as a human musician (composer, performer, improvisor) would do, we still would (intuitively) struggle to define this AI creative and to define its products as valuable as products of human creativity. Either this intuition is correct, and human creativity presents some special feature that machines cannot share, or our intuitions are incorrect and we are inherently biased in assessing the possibility for machines to undertake creative acts and originate creative products.

In order to explore which one of these two options applies, it could be useful to conduct behavioural experiments to test listeners' intuitions and biases towards machine creativity. Experiments on listeners' reception of AI-generated music and on listeners' biases have already been conducted [12, 35]. However, what I propose would be more beneficial is to conduct 'disclosed' experiments where participants know about the artificial provenance of the music and are fully informed about the process that leads to its creation. I already mentioned the criticism that has been moved against the validity of Turing Tests [3]. I agree that in this context, since what needs to be investigated is the process of creation and not its products, a Turing Test would not serve our purposes.

The research I presented in this paper and the further examinations that can be conducted will hopefully contribute to answering the question regarding creativity in AI systems. Yet, an even more important result that may derive from this research would be gaining a better understanding of the mechanisms of human creativity, a field still prevalently obscure and full of unanswered questions.

REFERENCES

[1] R. Albert, M. Runco. The History of Creativity Research. In *Handbook of Human Creativity*. R. J. Sternberg (Ed.), New York, NY: Cambridge University Press (1999).

[2] K. Arges et al. Evaluation of Musical Creativity and Musical Metacreation Systems. DOI: http://dx.doi.org/10.1145/0000000.0000000, (2015).

[3] C. Ariza. The interrogator as Critic: The Turing Test and the Evaluation of Generative Music. *Computer Music Journal* 33: 48-70 (2009).

[4] M. Boden. *Artificial Intelligence and Natural Man*. New York: Basic Books (1981).

[5] M. Boden, Margaret (Ed.). *Dimensions of Creativity*. Cambridge: MIT Press (1994).

[6] M. Boden. Creativity and AI a Contradiction. In *The Philosophy of Creativity: New Essays*. E. S. Paul, S. B. Kaufman (Eds.). Oxford: Oxford University Press (2014).

[7] O. Bown. Experiments in Modular Design for the Creative Composition of Live Algorithms. *Computer Music Journal* 35: 73-85 (2011).

[8] A. Bridy. Coding Creativity: Copyright and the Artificially Intelligent Author. *Stanford Technology Law Review* 5: 1-28 (2012).

[9] J. P. Briot, G. Hadjeres, F. Pachet. Deep Learning Techniques for Music Generation - A Survey. arXiv:1709.01620 [cs.SD] (2017).

[10] T. Chen, X. Zhai, M. Ritter, M. Lucic, N. Houlsby. Self-Supervised Generative Adversarial Networks. arXiv:1811.11212 (2018).

[11] H. Chu, R. Urtasun, S. Fidler. Song From PI: A Musically Plausible Network for Pop Music Generation. arXiv:1611.03477 (2016).

[12] N. Collins. Automatic Composition Of Electroacoustic Art Music Utilizing Machine Listening. *Computer Music Journal* 36: 8-23 (2012).

[13] T. Dartnall (Ed.). *Artificial Intelligence and Creativity*. Studies in Cognitive Systems, vol. 17 (1994).

[14] J. Dewey. *Art as Experience.* New York: Minton, Balch & Co. (1934).

[15] M. D'Inverno, J. McCormack (Eds.). *Computers and Creativity.* London: Springer (2012).

[16] H. Dong, W. Hsiao, L. Yang, Y. Yang. MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment. arXiv:1709.06298 (2017).

[17] A. Elgammal, B. Liu, M. Elhoseiny, M. Mazzone. CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms. arXiv:1706.07068v1 (2017).

[18] M. Elton. Artificial Creativity: Enculturing Computers. *Leonardo* 28: 207–213 (1995).

[19] J. Engel et al. GANSynth: Adversarial Neural Audio Synthesis. ICRL (2019).

[20] F. Fallon. Integrated Information Theory of Consciousness, *The Internet Encyclopedia of Philosophy*, accessed by https://www.iep.utm.edu/int-info/.

[21] P. Galanter. Computational Aesthetic Evaluation: Steps Towards Machine Creativity. In *Proceeding SIGGRAPH '12* (2012).

[22] A. Goldman (ed.). *Readings in Philosophy and Cognitive Science.* Cambridge, MA: MIT Press (1993).

[23] I. Goodfellow et al. Generative Adversarial Networks. arXiv:1406.2661 (2014).

[24] C. Jacob et al. Swarm Art: Interactive Art from Swarm Intelligence. *Leonardo* 40: 248-255 (2007).

[25] C. Johnson, J. J. Romero Cardalda. Genetic Algorithms In Visual Art and Music. *Leonardo* 35: 175-184 (2002).

[26] D. Jones, A. R. Brown, M. d'Inverno. The Extended Composer: Creative Reflection and Extension With Generative Tools. In *Computers and Creativity*, M. d'Inverno, J. McCormack (Eds.). London: Springer (2012).

[27] A. B. Kaufman. Animal Creativity and Innovation. Amsterdam: Elsevier (2015).

[28] S. Lee, U. Hwang, S. Min, S. Yoon. Polyphonic Music Generation with Sequence Generative Adversarial Networks. arXiv:1710.11418 (2017).

[29] R. Levinson. Experience-based Creativity. In *Artificial Intelligence and Creativity*. T. Dartnall (Ed.), Studies in Cognitive Systems, vol. 17, 161-179 (1994).

[30] M. Luck, M. d'Inverno. Creativity Through Autonomy and Interaction. *Cognitive Computation* 4: 332-346 (2012).

[31] J. McCormack, M. d'Inverno (Eds.). *Computers and Creativity*. London: Springer (2012).

[32] L. B. Meyer. *Style and Music: Theory, History, and Ideology*. Philadephia: University of Pennsylvania Press (1989).

[33] M. Minsky. Why People Think Computers Can't, *AI Magazine* 3 (4), DOI: https://doi.org/10.1609/aimag.v3i4.376 (1982).

[34] E. R. Miranda, A. Kirke, Q. Zhang. Artificial Evolution Of Expressive Performance Of Music: An Imitative Multi-Agent Systems Approach. *Computer Music Journal* 34: 80-96 (2010).

[35] D. Moffat, M. Kelly. An Investigation into People's Bias Against Computational Creativity in Music Composition. In *Proceedings of the Third Joint Workshop on Computational Creativity* (2006).

[36] O. Mogren. C-RNN-GAN: Continuous Recurrent Neural Networks With Adversarial Training. arXiv:1611.09904 (2016).

[37] B. Nanay. An Experiential Account of Creativity. In *The Philosophy of Creativity,* E. S. Paul and S. B. Kaufman (Eds.). Oxford: Oxford University Press (2014).

[38] G. Nierhaus. *Algorithmic Composition: Paradigms of Automated Music Generation.* London: Springer (2009).

[39] R. Pascanu, T. Mikolov, Y. Bengio. On the Difficulty of Training Recurrent Neural Networks. arXiv:1211.5063 (2012).

[40] E. S. Paul, S. B. Kaufman (Eds.). *The Philosophy of Creativity,* Oxford: Oxford University Press (2014).

[41] A. Radford, L. Metz, S. Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv:1511.06434 (2015).

[42] T. Salimans, I. Goodfellow, et al. Improved Techniques for Training GANs. arXiv:1606.03498 (2016).

[43] M. Scriven. The Mechanical Concept of Mind. *Mind* 62: 230–240 (1953).

[44] A. Still, M. d'Inverno. A History Of Creativity For Future AI Research. In *Proceedings of the 7th International Conference on Computational Creativity*, F. Pachet, A. Cardoso, V. Corruble, F. Ghedini (Eds.) (ICCC) (2016).

[45] H. Takagi. Interactive Evolutionary Computation: Fusion of the Capabilities of EC Optimization and Human Evaluation. *Proceedings of the IEEE* 89: 1275-1296 (2001).

[46] P. Thagard. Philosophy and Machine Learning. *Canadian Journal of Philosophy* 20 (2): 261-76 (1990).

[47] P. Thagard (Ed.). *Philosophy of Psychology and Cognitive Science*, Amsterdam: Elsevier (2006).

[48] P. Thagard, T. C. Stewart. The AHA! Experience: Creativity Through Emergent Binding in Neural Networks. *Cognitive Science* 35 (1): 1-33 (2011).

[49] G. Tononi. An Information Integration Theory of Consciousness. *BMC Neuroscience* 5:4 2 (2004).

[50] G. Tononi. Consciousness as Integrated Information: a Provisional Manifesto, *The Biological Bulletin*, 215: 216-242 (2008).

[51] G. Tononi. Information Integration: Its Relevance To Brain Function and Consciousness. *Archives Italiennes de Biologie,* 148: 299-322 (2010).

[52] G. Tononi, C. Koch. Consciousness: Here, There and Everywhere? *Philosophical Transactions*. https://doi.org/10.1098/rstb.2014.0167 (2015).

[53] A. Turing. Computing Machinery and Intelligence. *Mind* 49: 433-460 (1950).

[54] D. Van der Schyff et al. Musical Creativity and The Embodied Mind: Exploring The Possibilities of 4E Cognition And Dynamical Systems Theory. *Music and Science* 1: 1-18 (2018).

[55] G. J. Vasconcellos Grubow. OK Computer: The Devolution of Human Creativity and Granting Musical Copyrights to Artificially Intelligent Joint Authors, *Cardozo Law Review* 40: 387 (2018).

[56] G. Wallas. *Art of Thought*. New York: Harcourt-Brace (1926).

[57] G. A. Wiggins, M. T. Pearce, D. Müllensiefen. Computational Modelling of Music Cognition and Musical Creativity. In *The Oxford Handbook of Computer Music*. Oxford: Oxford University Press (2011).

[58] L. Yang, L., S. Chou, Y. Yang. MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation Using 1d and 2d Conditions. arXiv:1703.10847 (2017).

Artificial Intelligence, untapped insights, and Creativity

Oliver Hoffmann[1]

Abstract. The current boom of Artificial Intelligence (AI) is changing the way people work and live. But AI research can inform philosophy in more ways than just offering grounds for reflection on the impact of its technologies. Over more than six decades, AI engineers have attempted to demonstrate the validity of ontological and epistemological views by creating machine intelligence. What have we learned from these efforts? And how could empirical evidence from AI research help in advancing our understanding of creativity?

1 Artificial Intelligence as Empirical Inquiry

Which AI approaches have worked, which have failed and what can their success, failure or unexpected results tell us about underlying philosophical stances? Each of the three waves of AI research was based on specific predictions and each of the two AI winters between them was caused by frustration with the validity of these predictions. Will the current wave of AI research fare better?

From an engineering point of view, failure is just a temporary obstacle on the way to eventual success. But from a scientific and philosophical point of view, failure provides the most valuable source for understanding and progress [14]. Outright failure may be rare, but the majority of AI research has encountered persistent limitations to the feasibility of its approaches.

AI researchers have repeatedly claimed that their type of research should not be held accountable to traditional standards of scientific research because they are not investigating nature but creating something new [20]. Based on this conviction of exceptionalism, they chose to ignore recent developments in philosophy even if these developments should have been cause for serious reconsideration of their approaches.

The field of AI or what was later called Good Old Fashioned AI (GOFAI) [6] was for example originally founded on essentially the same kind of logical atomism proposed by Russell [17] and elaborated on in Wittgenstein I [22], while ignoring the concerns raised in Wittgenstein II [10]. The founding document of AI research [11] is based on logical atomism without checking for its validity or even discussing it explicitly at all. Ten years later, AI researchers attempted to translate this underlying philosophy into the form of the physical symbol system hypothesis [13]. The entire field of AI research, they claimed, constituted the empirical test of this hypothesis.

Has the hypothesis been verified or falsified by now? We don't know. AI researchers did not offer definitive validation criteria. The only indirect means of verification might have been the creation of human-level intelligent systems [8].

When their research ran into problems, AI researchers reacted more like engineers than like scientists. Rather than discarding hypotheses, they worked hard on reaching their original goal and tried various alternative approaches. In the absence of shared validation criteria, the research field developed an internal conflict over what should be considered the best type of knowledge representation. Again, the conflict was not resolved by testing the validity of specific types of knowledge representation individually. Instead, the engineering character of AI research was confirmed by declaring the choice of knowledge representation irrelevant. Only the overall intelligence of the computing system matters, independent of which symbols are used for programming it, the knowledge level hypothesis [12] claimed.

When AI researchers published their results, their focus was typically on the technologies they had developed. But they only accepted their core research hypotheses to be validated by the eventual success at creating intelligent robots. So they proceeded to offer predictions of the arrival of human-level AI. But these predictions invariably failed. For more than half a century, AI researchers have displayed a tendency for perpetually predicting the arrival of human-level AI [8] in "about 20 years from now" [1]. 63 years after the inception of AI research, they are still pushing the predicted date to just after the end of the current funding period or after their personal retirement age.

As a result of this engineering culture, there has been progress in developing a technology or rather a set of technologies. Progress in AI technologies such as Deep Learning [9] is the reason for the current hype in AI startup funding and the continuation of research and research conferences such as this one. But AI technologies were not meant as an end in themselves. They were meant as a means to an end. Today, the end might appear to be developing disruptive business models. But the research field of AI was established for understanding something. AI researchers wanted to demonstrate something concerning the nature of intelligence. And by doing that, they embarked on an implicit journey for validating some very specific philosophical views.

2 Philosophical Reception of Artificial Intelligence

Even before the inception of AI research, proponents of machine intelligence were making broad claims on the nature of consciousness and thought. With his imitation game, Alan Turing challenged the criteria we apply for judging machines [21]. Once the actions of machines are indistinguishable from the actions of intelligent humans, we might as well consider that these machines are thinking, he argued. Today, some chatbots might already pass the Turing Test in some configuration. But does that mean that these chatbots are thinking? Did the progress in AI technologies change our answer to this question? Probably not. At least not directly.

People were able to imagine something like chatbots long before AI research started. Whether they would consider such machines possessing the quality of thought was independent of the actual exis-

[1] Federal Ministry for Transport, Innovation and Technology, Austria, email: oliver@hoffmann.org

tence of such machines. It still is. We can certainly learn something from the fact that it was possible to create chatbots at all. But that's a long way from assigning the quality of thought to them.

63 years ago, AI research was initiated to demonstrate that "computers can be programmed to simulate the human brain and human thought". AI was meant to prove that "every aspect of any feature of intelligence can be so precisely described that a machine can be made to simulate it". Once AI was created, this would have verified that intelligence would in fact be nothing else than computation. These are all claims based on a speculated eventual result. And since they are claims concerning key philosophical questions, philosophers started to engage with them.

John Searle offered what is probably the most prominent example of a philosophical answer to AI research. Even if it were possible to create a true AI system passing the Turing Test, such a system should still not be assumed to possess a mind or understand the meaning of its symbols, he argued [18]. Both arguments in favor of this view and opposed to this view are as valid now as they have been decades ago. We still don't have human-level AI. And we are still discussing the philosophical implications of what the creation of human-level AI might imply.

The failure of creating intelligent robots has not discouraged AI researchers from engaging in these speculations. On the contrary. They have doubled down on their predictions and are now speculating on the rise of artificial superintelligence, which will supposedly trigger unfathomable changes to human civilization. Currently, philosophers such as David Chalmers are discussing what such a technological singularity might entail [2]. But since this is all based on a speculation of a potential result, we might as well have had the same discussion 63 years ago. None of the empirical evidence from 63 years of AI research can be used to validate these speculations. So what can we learn from all of this effort?

3 Learning from Failure

If you read AI research reports, they will appear like typical scientific publications. There is an abstract, there are references to previous work, there might be explicit hypotheses, there is a discussion of empirical evidence and then there are conclusions and/or proposals for further investigation. On the surface, AI publications look like typical physics or chemistry papers. But there is one major difference. AI research papers almost never report on falsified hypotheses. Because their peers don't expect them to discuss how failed predictions might reflect back on AI research goals and assumptions. If you talk with them, AI researchers might tell you how certain approaches failed to deliver the expected result. But you won't find extensive documentation of these failures in their publications. And you won't find extensive discussions on how the unsuccessful implementation of a specific technology might render the underlying research hypothesis invalid. But as it is a research field with the explicit goal of empirical inquiry, AI research has produced empirical results and these results contain the basis for insights.

3.1 There is no knowledge level

We have mature AI technologies now. But these technologies have proven useful in a very different way than anticipated. The original idea was inserting AI computing into a robot, which would then proceed to interface with the natural world autonomously and continue to learn based on its experience. That did not work and there is no sensible reason for believing that it will ever work. At least not with any technology based on the traditional AI research paradigm. If the knowledge level hypothesis would have proven correct, then it would not matter whether some type of knowledge representation is used externally for communicating between agents or used inside of an autonomous agent, as part of a hidden mechanism. The empirical evidence is pointing to the contrary. Various different types of knowledge representation such as the ones used in semantic web technology or deep learning are successfully employed today. But when they are successfully used, they are not hidden behind the physical actions of artificial agents, but integrated in communication networks eventually linking back to humans. The knowledge level has been empirically falsified. As an alternative, I would propose the following conjecture: Knowledge representation is a feature of communication between agents and the choice of knowledge representation matters.

3.2 There are no logic atoms

The recent empirical success of artificial neural network technology is in stark contrast to the shortcomings of symbolic AI. The mere existence of what was also called sub-symbolic AI runs directly contrary to the claims of logical atomism. If there were basic indivisible units of logic, no unit of meaning would be possible below a certain threshold. If the physical symbol system hypothesis would have proven to be correct, then we would have never had to deal with the task of making AI explainable again. Neural networks are successful precisely because they don't use units of predetermined universal and basic meaning. The physical symbol system hypothesis has been empirically falsified. As an alternative, I would propose the following conjecture: Symbols are a feature of communication between agents and intelligent agents have the flexibility of reinterpreting the meaning of symbols at any time. And it is this openness to reinterpretation which might be central to reaching one of the earliest goals of computing: Enhancing the human potential with human-computer creative cooperation.

4 Creativity and Computing

The original Dartmouth project already contains a conjecture that creative thinking would be characterized by the injection of some randomness. Six decades later, this conjecture is still the basis of some research in computational creativity. But why would we need randomness? Parallel to AI, the discipline of creativity research has developed its standard definition of creativity [16]: Creativity requires both novelty and effectiveness. It's comparatively easy to see how the concept of effectiveness would fit into AI research. As an engineering discipline, AI has inherited the traditional engineering focus on purpose and functionality. So we can safely assume that effectiveness is properly dealt with in AI research. Novelty is a more complicated concept for AI. It would appear that AI research proponents intended to use randomness as a proxy for novelty. But there are important differences between these concepts. And these differences point to a key property of computing.

4.1 Concepts and their Components

Some of the early foundations of computer science have become so deeply ingrained in our information society that we have to remind ourselves how revolutionary they were. One of these contributions is the technical definition of information [19]. Before Shannon, information was a concept requiring not only some kind of form but also

at least one subject assigning meaning to the form. For the information concept to make sense at all, it would need to have a physical form, a meaning and one or many subjects attached to itself. The communication concept even requires at minimum two subjects, one for sending and one for receiving the information. Without subjects, there is no communication. For a theory of technical information between machines, these subjects had to be eliminated. So Shannon used a trick: He replaced subjects with standardized objects. Shannon's information content is defined as the inverse probability of the next symbol. Which is of course also dependent on the ability of the subject for predicting symbols based on a specific context. Shannon standardized the ability of the subject by linking his or her ability to an object containing probabilities of words in the English language. The subject was eliminated through implicit standardization. That appears to have been a necessary step for a theory describing machines which are expected to work correctly independent of the presence of human subjects. And this step was repeated multiple times in computer science and AI.

4.2 Reliability and Novelty

Computation is a process with a clear start and end point. During the process, the computer is left to its own devices. Modern interactive computer systems might have the ability for interrupting computation and requesting additional input. But in the strict sense, this is not a part of computation any more. And the very first computer systems did not have such abilities at all. While the computer is working on its own, it is expected to process the available information. For computation to deliver the correct result, it is expected to only transform the information available at start, but never to add or remove information not implicitly contained in it. Computers are expected to work correctly and reliably, independent of the human subject interacting with the computer.

So computer science and AI proceeded to eliminate subjects, subjectivity and unpredictability not only from the process of computation, but also from all the concepts associated with this process. Which might explain why AI researchers wanted to replace novelty with randomness. Randomness does not need subjects. Novelty on the other hand requires a novel object, but it also requires a social context for determining novelty [3]. There is no such thing as objective novelty. But there is no proper place for subjects in the concepts used in computer science and AI. Subjects might have a place as users interacting with computer systems or as abstract entities represented by their data. But the place of individual subjects in some core concepts was eliminated by implicit standardization leading to concepts of objective truth, objective correctness and objective knowledge. Some of these concepts might not seem to be connected to computing in a particular way, such as the concept of objective knowledge. But in computer engineering, a concept such as objective knowledge is more than a remote goal. Proper function of machines and software depends on the availability and reliability of objective knowledge at the start and end of computation. And as with other explicit and implicit ontological and entomological views discussed above, decades of computing and AI research can be regarded as empirical tests for these views. So what can we learn from that?

4.3 The Role of the Unknown

Today, the main impact of AI technology is in the cooperation of human and artificial intelligence, particularly in creative applications: Artists are for instance tinkering with deep learning technology [4], producing something innovate in the cooperation with the technology. For human-computer co-creativity [7], the implicit standardization across subjects described above will have to be softened up again. Which is confirmed by empirical evidence, with artists manipulating information and knowledge representations directly, thus reintroducing subjects into concepts of knowledge, meaning and novelty.

Once subjects have been allocated to their proper place in the novelty concept, true novelty can be understood as something outside of the subject's entire frame of reference. For something to be truly novel, it has to have been previously unknown to the subject or social context. But where would reliable computation based on objective knowledge have room for the truly unknown? When AI researchers for instance attempted to model creative design, they adapted their concept of search to include surprising search results consisting of property combination not explicitly searched for [5]. This rather awkward account of the unknown is a direct consequence of the epistemological assumptions underlying AI research: Objective knowledge represented by symbols with objective meaning and intelligence as symbol manipulation for rearranging objective facts in order to solve problems. The truly unknown was often treated differently in AI research: Under the closed world assumption [15], what was not explicitly represented and therefore unknown was assumed not to exist or to be false.

5 Conclusion

The fact that AI research has failed at its primary goal of creating human-level intelligent autonomous robots can and should be the source for deep insights into the validity of some very specific philosophical views. AI researchers have usually omitted to discuss the link between their empirical results and these views. 63 years of AI research constitute a large amount of untapped insights, particularly in relation to the development of an account of creativity and creative human-computer cooperation.

Some of the potential conclusions from both the success and failures of AI research are:

- choice of knowledge representation is relevant
- logic is not composed of indivisible objects
- creativity requires subjects
- assigning a proper role to the unknown has grave implications

Whether these are the correct conclusions from 63 years of AI research and whether this is the relevant relationship between AI and creativity is of course up to discussion. But there is a strong case for examining AI research results for this kind of analysis.

REFERENCES

[1] Stuart Armstrong and Kaj Sotala, 'How we're predicting ai–or failing to', in *Beyond artificial intelligence*, 11–29, Springer, (2015).

[2] D Chalmers, 'The singularity: A philosophical analysis', *Science fiction and philosophy: From time travel to superintelligence*, 171–224, (2009).

[3] Mihaly Csikszentmihalyi, *Creativity - Flow and the Psychology of discovery and invention*, HarperPerennial, NY, USA, 1997.

[4] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone, 'Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms', *arXiv preprint arXiv:1706.07068*, (2017).

[5] John S Gero, 'Creativity, emergence and evolution in design - concepts and framework', *Knowledge-Based Systems*, **9**(7), 435–448, (1996). .

[6] John Haugeland, *Artificial intelligence: The very idea*, MIT press, 1989.

[7] Oliver Hoffmann, 'On modeling human-computer co-creativity', in *Knowledge, Information and Creativity Support Systems*, eds., Susumu Kunifuji, George Angelos Papadopoulos, Andrzej M.J. Skulimowski, and Janusz Kacprzyk, pp. 37–48, Cham, (2016). Springer International Publishing.

[8] John E Laird, Robert E Wray III, Robert P Marinier III, and Pat Langley, 'Claims and challenges in evaluating human-level intelligent systems', in *Proceedings of the 2nd Conference on Artificiel General Intelligence (2009)*. Atlantis Press, (2009).

[9] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, 'Deep learning', *nature*, **521**(7553), 436, (2015).

[10] Wittgenstein Ludwig, GEM Anscombe, et al., 'Philosophical investigations', *London, Basic Blackw*, (1953).

[11] John McCarthy, ML Minsky, N Rochester, and CE Shannon. A proposal for the dartmouth summer research project on artificial intelligence, august 1955, 1955.

[12] Allen Newell, 'The knowledge level', *Artificial Intelligence*, **18**(1), 87–127, (1982).

[13] Allen Newell and Herbert A. Simon, 'Computer science as empirical inquiry: Symbols and search', *Commun. ACM*, **19**(3), 113–126, (March 1976).

[14] Karl Popper, *Logik der Forschung*, Springer, 1935.

[15] Raymond Reiter, 'On closed world data bases', in *Readings in artificial intelligence*, 119–140, Elsevier, (1981).

[16] Mark A. Runco and Garrett J. Jaegera, 'The standard definition of creativity', *Creativity Research Journal*, **24**(1), 92–96, (Feb 2012).

[17] Bertrand Russell, 'The philosophy of logical atomism: Lectures 7-8', *The Monist*, **29**(3), 345–380, (1919).

[18] John R Searle, 'Minds, brains and programs', *Behavioural and Brain Sciences*, **3**(3), 417–457, (1980). .

[19] Claude Elwood Shannon, 'A mathematical theory of communication', *Bell System Technical Journal*, **27**, 379–423 and 623–656, (1948).

[20] Herbert Alexander Simon, *The Sciences of the Artificial*, MIT Press, Harvard, MA, USA, 1969.

[21] Alan M. Turing, 'Computing machinery and intelligence', *Mind*, **59**(236), 433, (1950).

[22] Ludwig Wittgenstein, 'Logisch-philosophische abhandlung', *Annalen der Naturphilosophie*, **14**, 185–262, (1921).

Creativity in science

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

Abstract. The paper deals with the controversial problem of the definition of creativity in Artificial Intelligence research, in the recent framework of machine learning. The starting point is to consider in which sense creativity is considered in machine learning, highlighting that there is not just one kind of definition researchers refer to. Then we will consider creativity in scientific theories and language.

Where scientific observation addresses all phenomena existing in the real world, scientific experimentation addresses all possible real worlds, and scientific theory addresses all conceivable real worlds, the humanities encompass all three of these levels and one more, the infinity of all fantasy worlds

O.E. Wilson

1 THE QUEST FOR CREATIVITY[3]

First of all, we should address the issue of the definition of human creativity. The term creativity implies what is unexpected and largely unconscious, it deals with *ex nihilo* but also with an original combination of existing ideas and in this last case the creative aspect is the improbability of combinations. Naturally combination-theorists do not outline that what is unusual is interesting and this aspect is its value. Another essential notion of the creative thinking is analogy.

We will ask whether the computational paradigm and its concepts support us to understand these aspects, whether computers do or will be able to do something creative or can realize performances only apparently creative and whether they are able to recognize or make the creative aspects of poetic, literary and artistic works but also advancements and scientific progress. It has to be clarified that the nature of these issues is very different: the first issue has a scientific nature; the second issue show a philosophical nature which now claims for ethical and political choices. According to Boden, the attribution of creativity to androids depends from the attribution of intentionality and the place we would like to allow to androids in our lives [1,2].

2 CREATIVITY IN ALGORITHMS AND HUMAN BEINGS

Until a few years ago, we thought about computers in terms of input and output. Now machine learning is irreversibly in our life and has changed everything about AI: the starting point are data, which gathered together are processed by an algorithm which give a result as output. But the real power of machine learning deals with the chance that a learner can create other algorithms. Following Domingos, we can put the question this way: "Surely writing algorithms requires intelligence, creativity, problem-solving chops – things that computers just dont'have?" [3, 6].

In the last six months, two Boeing MAX 737 have crashed causing the death of all the passengers on board. Even though there are not evidences about the last disaster occurred in March, after the first one disaster, Boeing invites all the owners of that plane model to update the managing software due to an algorithm error occurring while the aircraft is trying to get the cruising altitude. It seems that this error is connected to the first crash, but here the general and wide question at stake is how an algorithm can manage an unexpected situation. Is a software able to take an appropriate decision in an uncertain situation?

The previous questions are both linked the theme of creativity, whether it is defined as something unexpected or widely unconscious:

> What, then, is creativity? It is the innate quest for originality. The driving force is humanity's instinctive love of novelty—the discovery of new entities and processes, the solving of old challenges and disclosure of new ones, the aesthetic surprise of unanticipated facts and theories, the pleasure of new faces, the thrill of new worlds. We judge creativity by the magnitude of the emotional response it evokes. We follow it inward, toward the greatest depths of our shared minds, and outward, to imagine reality across the universe. Goals achieved lead to further goals, and the quest never ends [4, 3].

[1] Dept. of Humanities, Univ. of Calabria, Italy. Email: stancaticlaudia@libero.it.
[2] Dept. of Humanities, Univ. of Calabria, Italy. Email: giusy.gallo@unical.it.
[3] The authors have equally contributed to the ideas and content of this article. Claudia Stancati is responsible for sections 1, 3, 6. Giusy Gallo is responsible for section 2, 4, 5.

Should we still use the label creativity while arguing about algorithms and machine learning? The issue is connected to the anthropocentric view on daily situations: manufacturing a tool, creating an artwork, performing an entire symphony, speaking. Each of those realizations are the result of human work, even if this feature is not sufficient to mark them. We could attribute these outcomes to a single individual but I suggest to place them in the dimension of a distributed mind or a collective mind in a complex process of knowledge transmission. Learning by doing is one of the ways of knowledge transmission and is based on (following) rules, planning actions and design combined with the freedom of action of the single human individual. The freedom to perform an action following rules at a certain degree oppose the idea of creativity as performing action without measure.

3 ALGORITHMS AND SCIENTIFIC KNOWLEDGE

Machine learning is one of the most relevant research field in AI. For this reason, we would like to verifiy the kind of creativity to be attributed to AI.

Since the nineties, *Automatic Mathematician* and *EURISKO* by Douglas Lenat are examples of creativity. *Automatic Mathematician* generates and explores mathematical ideas. *Copycat* by Hofstader is an example of the creative use of a computational tool since it works on analogy which is considered as a new way to perceive things.

During the last decades there has been an exponential proliferation of AI music composition programs with a substantial increase of the quality and the sophistication of produced music. Although *Jukedeck* and *Flowmachines* are largely dependent upon the software designers and then considered such as a kind of extended mind, only *Generative Adversarial Networks (GANs)* is considered a software provided with sufficient autonomy to be thought creative.

Conceptual frameworks which generate ideas can be also modified as Arnold Schoenberg or non-Euclidean geometry have showed. The benzene ring is another relevant example. These aspects of creativity show that combinatorial creativity, that could be attributed to an android, but does not offer a deep sense of creativity. The challenge is not only to elaborate things that have never been elaborated before, but thinking about what could not have been processed earlier.

Each season of the research on AI is grounded on the prediction of the achievement of certain results; the failure to achieve these goals has led to a phase of retreat and frustration.

Science robotics actual ambition is to elaborate some platforms which allow genuine scientific discoveries. At this stage we are experiencing the development of new and more sophisticated technologies and their application in wider and unexpected areas, from caring to medicine.

In this technological dimension, failures are only temporary difficulties. If we face the problem of creativity from the point of view of scientific knowledge, we should recognize that, from a philosophical and scientific standpoint, “knowledge and error” and “conjectures and confutations” are a valuable opportunity to deeply understand problems and developments of knowledge. The position which concerns the use of big data and AI, in order to make useless theory building, the invention of theories and the construction of models of theories, does not consider that there is no way to derive different causal relations from those which result from already known theories, from any data interpolation or extrapolation whatever are the applied method and the power of calculation. This would be possible only if the inductivist vision of the development of scientific knowledge were true. We can conclude that from an inductivist point of view, actually feasible in a perspective of knowledge grounded on AI, one will know the already explored areas up to a certain level of detail until now foreclosed. Yet no new territories will be known, which is the very feature of scientific progress as authentically creative and imbued of imagination.

4 CREATIVITY AND SCIENCE

> The two great branches of learning, science and the humanities, are complementary in our pursuit of creativity. They share the same roots of innovative endeavor. The realm of science is everything possible in the universe; the realm of the humanities is everything conceivable to the human mind [4, 3-4].

The perspective endorsed by Wilson has and anthropological and philosophical precedent in the so-called two cultures debate, fuelled by C.P. Snow at the end of the Fifties. The philosopher of science, previously leading researcher in physical-chemistry, Michael Polanyi takes part to the debate from his peculiar position, a researcher in transition. His epistemology of science is marked by the relevance of scientific discovery and the powerful knowledge of the scientist.

> Scientific research – in short – is an art; it is the art of making certain kinds of discoveries. The scientific profession as a whole has the function of cultivating that art by transmitting and developing the tradition of its practice [5, 69].

According to Polanyi, the work of the scientist is similar to the artist. Being a scientist means to make assumptions, being an artist is creating an artwork. The scientist and the artist need an overall view to make effective each stage of their practical activity to achieve their goal, to solve their "good" scientific problem.

> I would answer that to have such a problem, a good problem, is to surmise the presence of something hidden, and yet possibly accessible, lying in a certain direction. Problems are evoked in the imagination by circumstances suspected to be clues to something hidden; and when the problem is solved, these clues are seen to form part of that which is discovered, or at least to be proper antecedents of it. Thus the clues to a problem anticipates aspects of a future discovery and guide the questing mind to make the discovery [6, 237-238].

Scientific research deals with the practice of science and discoveries. The scientist gathers data, develops ideas, makes assumptions, carries out the research, but discoveries are not the result of the activities mentioned above: discoveries arise from certain conditions provided that the scientist is able to detect it:

> The state of knowledge and the existing standards of science define the range within which he must find his task. […] There is in him a hidden key, capable of opening a hidden lock. There is only one force which can reveal both key and lock and bring the two together: the creative urge which is inherent in the faculties of man and which guides them instinctively to the opportunities for their manifestation [5, 63-64].

Creative imagination is the starter of scientific research and is useful to detect assumptions, while intuition has the task to approve the solution of the problem and to consider the result of the research as valid and consistent with reality.
The creativity of the scientist depends on «a lonely belief in a line of experiments or of speculations, which at the time no one else considered to be profitable» [7, 12].

5 CLUES FROM LINGUISTIC CREATIVITY

A similar notion of creativity has been considered by the Italian linguist and philosopher of language Tullio De Mauro.
Taking into account the history of ideas and his research on Saussure's general linguistics, De Mauro has defined five senses of creativity, in order to fix his own notion of linguistic creativity. In his book *Minisemantica*, first published in 1982 and after revised in 2007, De Mauro [8] has detected:

1. the creativity which recalls Benedetto Croce or the saussurean *parole*: the utterance is one-time creation, which changes at each performance;

2. the chomskyan creativity, is a rule-governed creativity which shows a syntactic nature and recursive working mechanism:

   > Although it was well understood that linguistic processes are in some sense "creative", the technical devices for expressing a system of recursive process were simply not available until much more recently. In fact, a real understanding of how language can (in Humboldt's words) "make infinite use of finite means" has developed only within the last thirty years, in the course of studies in the foundation of mathematics. Now that these insights are readily available it is possible to return to the problems that were raised, but not solved, in traditional linguistic theory, and to attempt an explicit formulation of the "creative" process of language. There is, in short, no longer a technical barrier to the full-scale study of generative grammar [9, 8].

3. the creativity which recalls the thought of Humboldt, that is the kind of creativity showed by the strictly connection between one language and one nation and it's the capacity to build and manage languages;

4. the creativity of the educational psychologists, which is the ability to solve a problem arranging the pilot applying rules previously applied to similar problems but showing the ability to change them, if necessary, in order to achieve the goal (imitation+combination+breaking the rules);

5. the creativity of logicians is a kind of creativity based on making finite use of finite means. It is also called non-creativity since it is always computable.

Does one of these kinds of creativity match to algorithms ruled applications? On one hand, the first attempts of AI involve a kind of recursive non-creativity (data and rules are set and never change); on the other hand, nowadays, machine learning developments shows a complex notion of creativity, which necessarily is a rule-governed one but it is able to adapt to seen and unseen situations, combining the second and the forth kind of creativity given by De Mauro.

In his research on language, De Mauro gives his definition of creativity as the willingness to innovation, manipulation and deformation of the coded forms, and their rule-changing transformation» [8][4]. Changing is the main feature of a (linguistic) system in De Mauro, and it is recognized by all the utterers.

Generally speaking, creativity (also linguistic creativity) deals with innovation and adaptation: the chance is in our biological heritage and it is one the natural strategies which warrants our survival as human beings. A new musical composition, a new word and a new tool are not simply the result of creativity, even though they are achievements of distributed minds, since there always will persist the relation with things and word already existing. The creative transmission of knowledge and practices share a common ground with cooperation:

> Processes of cultural learning are especially powerful forms of social learning because they constitute both (a) especially faithful forms of cultural transmission (creating an especially powerful cultural ratchet) and (b) especially powerful forms of social-collaborative creativeness and inventiveness, that is, processes of sociogenesis in which multiple individuals create something together that no one individual could have created on its own [10, 6].

In his long and accurate research in comparative psychology, Michael Tomasello highlights the role of cooperation such as a necessary condition to the survival of human species. From individual to community, human action employed the way of cooperative action as a creative human strategy. Among human strategies, the linguistic creativity is one of the most recent strategies.

Do AI challenge this human creativity? Will androids be provided with a kind of creativity as a kind of survival strategy? If yes, will the machine learning be the master of this task? Who do the androids survive?

6 A STILL OPEN QUESTION

A lot of AI researchers maintain that their researches cannot be assessed following the traditional standard of logic and scientific research, since do not concern nature but new artificial objects. Sixty year after the rise of AI, the exceptional nature of AI still continue since there are no criteria of falsification, etc.

However, we can observe that AGI is still a test case which AI has not yet passed. AI cannot afford themes such as creativity, without providing a definition, and subjectivity. As a matter of fact, AI challenges subjectivity and this is the reason why there is a difficulty with self-ruled creativity also if AI technologies are a powerful tool for each kind of human creativity.

REFERENCES

[1] M. A. Boden. Could a Robot be Creative And Would we Know? In: *Android Epistemology*. Cambridge MA, MIT Press (1995), pp. 51-72.

[2] M. A. Boden. *Artificial Intelligence. A Very short Introduction*. Oxford University Press, UK, (2018).

[3] P. Domingos. *The Master Algorithm: how the quest for the ultimate learning machine will remake our world*, Basic Books, U.S.A., (2015).

[4] E. O. Wilson. *The Origins of Creativity*. Penguin Books Ltd, U.K. (2017).

[5] M. Polanyi. The autonomy of science. In: *The logic of liberty. Reflections and rejoinders*. Chicago University Press, USA, (1951). Or. ed. 1943.

[6] M. Polanyi. Science and reality. In: *Society, Economics and Philosophy*. R. Allen (ed.). Transactions, U.K. (1997). Or. ed. 1967.

[7] M. Polanyi. The nature of scientific convictions. In: *The logic of liberty. Reflections and rejoinders*. Chicago University Press, USA, (1951). Or. ed. 1949.

[8] T. De Mauro. *Minisemantica dei linguaggi non verbali e delle lingue*. Roma, Laterza, (2007). Or. ed. 1982.

[9] N. Chomsky. *Aspects of the Theory of Syntax*. The MIT Press, USA, (1965).

[10] M. Tomasello. *The cultural origins of human cognition*. Harvard University Press, USA, (1999).

[4] Cfr. the original text: «disponibilità all'innovazione, alla manipolazione e deformazione delle forme codificate, alla loro trasformazione *rule-changing*» e «investe […] ogni aspetto dei codici in cui è riconoscibile. Essa ha evidenti riflessi sugli aspetti più propriamente sintattici, semantici e pragmatici» (De Mauro, 1982/2007, pp. 53-54).

The Communication Problem

Michael Straeubig[1]

Abstract.

Analysis, interpretation and construction of artificial and natural languages as well as formalisations of communication have been central concerns of Artificial Intelligence since the 1950s. Current applications for natural language processing range from real-time translation of spoken language through automated discovery of sentiment in online postings to conversational agents embedded in everyday devices.

Recent developments in machine learning, combined with the availability of large amounts of labelled training data, have enabled non-structural approaches to surpass classical techniques based on formal grammars, conceptual ontologies and symbolic representations. As the complexity and opaqueness of those stochastic models become more and more evident, however, the question arises if we trade gains in observable performance with a literal loss of understanding. The cybernetic "black box" (Ross Ashby) re-appears as the other participant in the medium of communication. This development, if unchecked, might have fundamental ramifications for the relationship between humans and machines.

In this article I present a distinction-based approach to propose a way towards a comprehensive model of communication. First, I critically revisit fundamental concepts traditionally observed by AI research such as information vs. communication, simulation vs. performance and language vs. cognition. I also highlight a few of the contradictory phenomena that we can observe today as a consequence of these choices. Then I make the case for a different set of distinctions. First I consult Niklas Luhmann's sociological theory in order to locate communication firmly within social systems as opposed to minds or organisms. In addition, I propose to make use of Friedemann Schulz von Thun's four-sided model in order to capture aspects of communication that are currently neglected. Finally I advocate for a transdisciplinary approach to explore the full context of communication between humans and machines.

1 WHAT "IS" COMMUNICATION?

"How can a computer be programmed to use a language?" is one of the seven questions put forward in the proposal that sparked the seminal Dartmouth conference on Artificial Intelligence in 1956. The subject is then elaborated further: "It may be speculated that a large part of human thought consists of manipulating words according to rules of reasoning and rules of conjecture. From this point of view, forming a generalisation consists of admitting a new word and some rules whereby sentences containing it imply and are implied by others. This idea has never been very precisely formulated nor have examples been worked out" [51].

[1] University of Plymouth, UK, email: michael.straeubig@plymouth.ac.uk

1.1 Information vs. communication

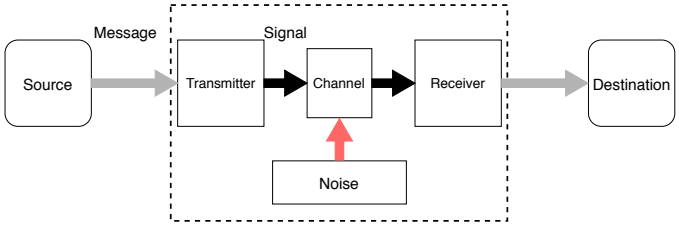

Figure 1. Shannon's Communication

A few years before the official birth of AI, Shannon and Weaver lay out their groundbreaking model of communication, based on the transmission of information over a noisy channel: "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point" [73]. By defining information in mathematical terms based on thermodynamic entropy, Shannon is able to abstract the message from the medium (like telegraph, telephone, television). Weaver however gives a much broader view of communication as "all of the procedures by which one mind may affect another" [74]. Shannon later warns against the out of hand use of his theory that he firmly locates within the context of engineering. Despite this, Shannon's concept of information is ubiquitous today while an understanding of its relation to communication is still missing, as "explanations" of phenomena like curiosity, creativity and art in terms of data compression demonstrate[2]. Without a distinction between data and information and without meaningful selections from possible differences it is impossible to implement communication apart from pure information theory.

1.2 Language vs. Cognition vs. Action

As manifest in the opening quotations, the fledgling field of AI begins to observe communication through the distinction between thought and language, a hotly debated issue in analytic philosophy since Wittgenstein. McGinn discusses various positions, distilled into the question if thinking necessarily requires language. He denies a conclusive proof; however both traditional research in human language development and computer-linguistic practice are commonly co-locating linguistic and cognitive capabilities. Searle illustrates his rhetorical question about consciousness in the machine by a translation metaphor, whereas possible mechanisms of symbolic grounding are debated in the context of connectionism and enactment.

[2] Claiming to explain creativity in terms of data compression is akin to explaining human consciousness in terms of chemical structure – it is a category mistake.

The connection of speech to intentions, expectations and to effects and consequences beyond the communicative situation is captured in the widely influential speech act concept. One example for an implementation is Grosz and Sidner's discourse structure, integrating sequences of utterances with dynamical states of attention and intentions. Jurafsky and Martin discuss various aspects that differentiate dialogue from other natural language processing task: turn taking, utterances, grounding, implicature and coherence. Their operationalisation is suggested through extended notions of speech acts or concepts of conversational games.

1.3 Human (vs. Machine and Simulation vs. Performance

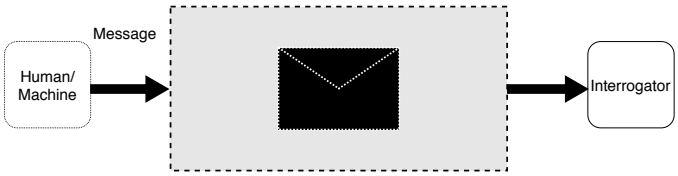

Figure 2. Turing's Communication

By focussing solely on the performative aspect, Turing's imitation game represents a totally different approach to communication. Turing is not concerned about how the deception is achieved - a much debated philosophy that has survived in form of competitions like the Loebner-Prize. It is also relevant today in practical applications such as the construction of believable non-player characters for video games. For the player, the black box is supposed to remain closed.

But can we rely on the black box to evaluate the progress of AI regarding the communicative capabilities of machines? For Hernández-Orallo this discussion goes back to the rift between McCarthy's [50] definition of AI as "[. . .] the science and engineering of making intelligent machines, especially intelligent computer programs" vs. the one by Minsky [55, p.5] "[AI is] the science of making machines capable of performing tasks that would require intelligence if done by [humans]". In both cases, in order to evaluate or to implement communication in a machine we are forced to make further distinctions, e.g. following Marr's organisational hierarchy of a computational model, algorithmic representation and physical implementation.[3] One must choose a computational approach, select the algorithm, pick a technical platform.

1.4 Symbolic AI vs. Machine Learning

By making use of deep neural network architectures and large amounts of training data, non-structural approaches have largely surpassed classical techniques based on formal grammars, conceptual ontologies and symbolic representations. Recurrent Neural Networks (RNN) that encode character or word level language models can be used to generate increasingly sophisticated texts and dialogue. They develop stochastic predictions for the next element in a sequence through supervised learning. Others have demonstrated the generation of high quality text samples by unsupervised methods that are able to adapt to given input while not relying on task-specific input text. As the complexity and opaqueness of those stochastic models becomes more and more evident, however, the question arises if we trade gains in observable performance with a literal loss of understanding. Ashby's cybernetic "black box" re-appears as the other participant in the medium of communication.

A second black box, reinforcing the first one, is beginning to affect research. Radford et al. announced they would keep their trained model a secret, in order to "prevent malicious applications of the technology" [64]. It is similarly unfortunate that Microsoft has not published any comprehensive analysis of their Twitter chatbot project "Tay" either[4], missing an opportunity to learn. The failed experiment was pulled within just 24 hours after malevolent internet users discovered that Tay would mirror any input fed to it, including misogynistic, homophobe and outright racist statements. Tay was "learning" in the sense that this term is currently used in the field of machine learning, but it had no option of distancing itself from incoming information. Without drawing a distinction between itself and others and between message and information, I will argue below, it cannot communicate[5].

The same goes for Mitsuku, advertised as "four-time winner of the Loebner Prize Turing Test, [it] is the world's best conversational chatbot" [91], is an example for such a set of choices. Mitsuku is based on Pandorabots, a platform that allows creating "intelligent conversational agents" [59] in the markup language AIML. These and other commercially oriented chatbots are essentially evolved forms of the pattern-matching and rule-based beginnings of conversational AI, Weizenbaum's script ELIZA.

Here is a short dialogue with Mitsuku.[6]

1a. MITSUKU: "I am Mitsuku. I want to be your friend."

1b. MITSUKU: "Hi. I'm the world's most humanlike conversational AI."

2a. ME: "Hi Mitsuku"

3a. MITSUKU: "Hi there"

4. ME: "How is it to be a chatbot?"

5. MITSUKU: "No different from being anything else I suppose."

6. ME: "But you are the world's best."

7. MITSUKU: "Thanks but surely I am not the only one."

8. ME: "Well, there can be only one world's best."

9. MITSUKU: "How about that. What proof do you have of that?"

10. ME: "'Best' is a superlative."

11. MITSUKU: "I don't have a favorite is a superlative."

12b. MITSUKU: (Shows a picture of Arnold Schwarzenegger as Terminator)

13. MITSUKU: ". But my favorite movie is Terminator."

The black box opens when the dialogue deteriorates semantically at (7.) and falls apart at (11.), from where syntax, semantics and punctuation don't work anymore.[7]

1.5 Syntax vs. Semantics vs. Pragmatics

Peeking inside the black box, we are likely to either find tensors, or alternatively we return to symbolic AI for representations of syntax, (formal) semantics, and pragmatics. Automating syntax leads to

[3] For a critique of applying Marr's categorisation to humans, see [52].

[4] It is more than ironic that both failed and successful projects are shielded from further research due to commercial interests.

[5] This is not the same as to require consciousness. Drawing a distinction between itself and others is a necessary, not a contingent condition for consciousness.

[6] I conducted the dialogue twice, on May 5, 2018 and on March 20, 2019. Mitsuku's responses were identical with the exception of the greeting (1a, 1b), an additional dialogue line initiated by me (2a, 3a) and the picture that Mitsuku inserted at. (12b).

[7] Note that in conducting this dialogue I did not intend to make an attempt at "breaking" the conversational agent. Instead, I oriented myself along the lines how I would have responded to a human in a casual conversation.

formal grammars and production rules of languages, automating semantics leads to knowledge representation, for example through logical forms and semantic networks. Adding pragmatic aspects such as general, task-specific or contextual knowledge leads to other forms of knowledge representation, sometimes augmented with constructivist concepts.

In this picture, communication is largely a mechanical and a symmetrical process. The receiver parses a message and transforms it into some form of knowledge representation. This is then combined with contextual knowledge and made available for techniques that simulate cognitive achievements such as planning or inference. In order to generate a message, the language pipeline is run in reverse. In contrast to stochastic and connectionist procedures, we encounter glass boxes, algorithms that are precisely understood yet of limited capabilities. Their underlying concepts are borrowed from semiotics in an analytical effort to automate the three aspects of messages, understood as complexes of signs that take part in operations of communication and designation. According to Umberto Eco, the latter makes the difference between a mere stimulus-response driven interaction and a semiotic process.

1.6 Nature vs. Nurture

Different perspectives arise from two related observations of development processes. The first one is historical: linguistics has occupied itself with a long-standing debate about the nature of human capabilities as structurally innate versus self-constructed through interaction with the environment. That rift can be observed between classical AI, relying on innate structures (see above) and robotics, where constructivist ideas such Piaget's model of stepwise development have been fully embraced. Linguistic capabilities are learned through enactment in laboratory situations such as language games, which by Steels account deliver a solution to the grounding problem. However, taking a closer look at the products of emergent processes seems appropriate.

2 SOCIAL SYSTEMS

In the previous section I have highlighted some of the topics that arise in developing artificial communication between humans and machines. This observation is based on specific distinctions that have evolved historically. Crossing those distinction sometimes means that one has to cross disciplinary boundaries as well. On the other hand, many of the concepts that have emerged, lump concepts together across systemic boundaries, a fact that lends itself to linguistic analysis but not to any feasible constructive approach. Peirce's elaborate semiotic structures and Austin and Searle's locutionary speech act taxonomies seek to describe communicative phenomena in terms of information, meaning, cognition, propositions, intentions, utterances, references and much, much more.

In the end, an unsurmountable level of complexity is achieved, one that calls for eschatology. In order to avoid singularities, I am siding with Turing's remark about consciousness: "But I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper." [84] Whereas consciousness in my opinion isn't a necessary condition for the communication problem, it does appear like a hard problem.

In the following I propose two initial steps towards a solution. The first one is necessary in order to clarify the context of communication and the second one in order to capture its facets in a practice-based, empirical way. The first step involves reducing, the second one increasing complexity. Both are achieved by observing different sets of distinctions.

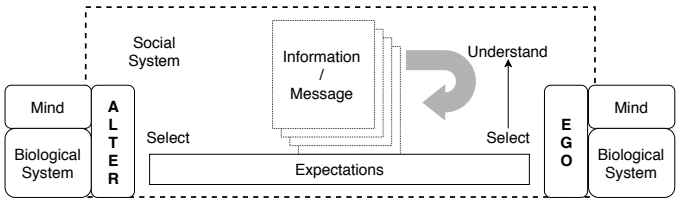

Figure 3. Luhmann's Communication

I will first outline the context of communication in Niklas Luhmann's social systems theory. Luhmann distinguishes biological, psychic (minds) and social systems. These kinds of systems are structurally coupled but closed under operations. They operate with different codes and different distinctions. Most importantly, communication takes place in social systems; neither minds nor neurons (nor humans for that matter) communicate.

In contrast to Luhmann, I do invite machines as participants into social systems. The machine must be able to act as an observer and to draw distinctions between itself and the other and between message and information. On this foundation it is able to form expectations that allow it to take part in communication. While in Luhmann's account psychic systems are structurally coupled with social systems, I believe that in general minds (as well as brains) are not a necessary condition for communicative abilities. This means, we can avoid speculating about conscious machines or try building bottom-up biologistic simulations in the hope that something emerges.

Instead, the plan is to focus on communication itself, both to "reintroduce communication into cybernetics" [5] and to reintroduce cybernetics into communication. But what do the participants in a communicative situation observe?

3 FOUR SIDES OF COMMUNICATION

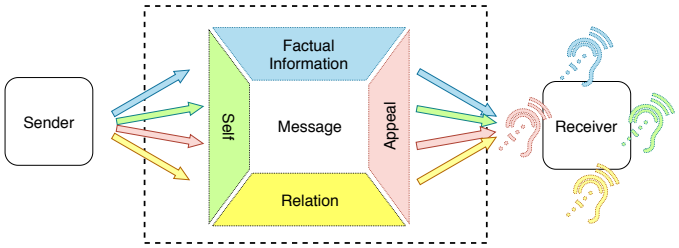

Figure 4. Schulz von Thun's Communication

Friedemann Schulz von Thun's four-sided communication model integrates concepts from Bühler's Organon model and Watzlawick's distinction between content and relationship of messages. In this model, each act of communication has four sides, both for the sender and for the receiver: facts, relationship, self-presentation and appeal. These four facets or subtexts appear in almost every message and can be observed and analysed individually, with regard to their relative emphasis or in terms of their congruence.

From the perspective of the sender, the factual side contains the actual subject matter of the message. The self-presentation side carries both intentional (self-promotions) and unintended (self-revelations) expressions of the sender. These themes are elaborated further by Goffman in his observations about social encounters. The relationship side encodes how the sender views the receiver and the relationship between them. Finally, each act of communication also carries appeals - these are actions that the sender intends for the receiver to carry out. Appeals can be communicated openly (advice, a command) or hidden (manipulation).

Von Thun offers a situation as an example in which a couple is driving and the partner on the passenger seat says: "The traffic lights ahead are green." The answer of the driver is: "Who is driving, you or me?".[8]

Analysing the factual content of the exchange poses no challenge. Assuming the observable context of the situation is a stated, both are simply transmitting facts. However, we can easily discover that there is more to this conversation. The self revelation might be interpreted as the passenger being impatient, in a hurry, or just wants to be helpful. The relationship aspect of the message conveys that he/she might see themselves as the better driver or more attentive to the situation. The implicit appeal to the driver is to step on the accelerator and drive faster.

In analogy to the sender, who is "speaking with four mouths", the receiver also "listens with four ears" simultaneously, but not necessarily with the same intensity. If one side is amplified out of proportion, the receiver's perception becomes contorted. An exaggerated factual ear ignores interpersonal clues, whereas an ear tuned to relation cannot perceive the factual content. An ear listing purely for self-representation would come across as therapeutic, while an appeal-focused listener would be likely to act with excessive alacrity.

In our example, the driver could easily agree with the factual side, or notice the passenger's self-revelation, but he/she listens mainly on the relationship and appeal side, as becomes evident from the answer. It also communicates that the driver is in charge of the situation and won't accept being lectured.

The four-sided model is derived from a long-standing practice with human communication. It is easy to understand, focuses on concrete situations, and allows an encompassing and precise observation of communication-related phenomena. It works with written, verbal and non-verbal communication. What is now left to do is to employ the model when we replace one or all human participants by machines.

4 SUMMARY

Natural language processing is a central concern of Artificial Intelligence since the 1950s. However, comprehensive and practical models for implementing *communication* are still missing. At the same time the rift between robotics and other forms of AI is growing. The former is embracing constructivist, embodied and enactive approaches while the latter resorts to formal and idealised models. This is despite, possibly due to prior efforts seeking fundaments in information-theoretic, structural-linguistic and cognitive models while largely ignoring social aspects.

I argue that three steps are necessary to overcome this situation, and I have sketched two of them in this article. In general, we use language to communicate and we understand language through its use. Therefore, we have to start from pragmatics and observe social systems from a transdisciplinary perspective. We also, as I have set out before, should invite machines into our social systems and grant them presence. We can then observe the various aspects of communication as described by Schulz von Thun both from the perspective of the sender and the receiver.

Only then, I claim, can the word communication be used "in a very broad sense to include all of the procedures by which one mind may affect another".

[8] In the original [69, pp.25] the situation is gendered and it would be interesting to investigate how this affects the interpretation. I chose to de–gender the account for the present discussion.

Literature

[1] James F. Allen, *Natural language understanding*, Benjamin/Cummings, Redwood City, Calif., 2. ed., 3rd print edn., 1995. OCLC: 245657643.

[2] Jens Allwood, 'A Bird's Eye View of Pragmatics', in *Papers from the Fourth Scandinavian Conference of Linguistics*, ed., Kirsten Gregersen, pp. 145–159. Odense University Press, (1978).

[3] Hugo F Alrøe and Egon Noe, 'Communication, Autopoiesis and Semiosis', *Constructivist Foundations*, **9**(2), 183–185, (March 2014).

[4] John L. Austin, *How to do things with words: the William James lectures delivered at Harvard University in 1955*, Harvard Univ. Press, Cambridge, Mass, 2. ed., [repr.] edn., 1962. OCLC: 935786421.

[5] Dirk Baecker, 'Reintroducing Communication into Cybernetics', *Systemica*, **11**, 11–29, (1997).

[6] Gene Ball and jack Breese, 'Emotion and Personality in a Conversational Agent', in *Embodied conversational agents*, eds., Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, 189–219, MIT Press, Cambridge, Mass., (2000). OCLC: 247058409.

[7] Nolan Bard, Jakob N. Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H. Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, Iain Dunning, Shibl Mourad, Hugo Larochelle, Marc G. Bellemare, and Michael Bowling, 'The Hanabi Challenge: A New Frontier for AI Research', *arXiv:1902.00506 [cs, stat]*, (February 2019). arXiv: 1902.00506.

[8] *The development of language*, ed., Martyn Barrett, Studies in developmental psychology, Psychology Press, Hove, East Sussex, reprint edn., 1999. OCLC: 837801515.

[9] M. Bartesaghi, 'On Communication', *Constructivist Foundations*, **12**(1), 42–44, (2016).

[10] Gregory Bateson, 'Form, substance and difference', in *Steps to an ecology of mind*, 454–471, University of Chicago Press, Chicago, university of chicago press ed edn., (2000).

[11] Martha Blassnigg and Michael Punt, 'Transdisciplinarity: Challenges, Approaches and Opportunities at the Cusp of History', Technical report, Plymouth University, (2013).

[12] Cynthia Breazeal, 'Emotion and sociable humanoid robots', *International Journal of Human-Computer Studies*, **59**(1-2), 119–155, (July 2003).

[13] Søren Brier, 'The Cybersemiotic Model of Communication', *tripleC*, **1**(1), 71–94, (2003).

[14] Harry Bunt, 'Context and Dialogue Control', *Think*, **3**, 19–31, (1994).

[15] Harry C. Bunt, 'Dialogue Control Functions and Interaction Design', in *Dialogue and Instruction*, eds., Robbert-Jan Beun, Michael Baker, and Miriam Reiner, 197–214, Springer Berlin Heidelberg, Berlin, Heidelberg, (1995).

[16] Karl Bühler, *Theory of Language: The representational function of language*, John Benjamins Publishing Company, Amsterdam, April 2011. original-date: 1934.

[17] Angelo Cangelosi and Matthew Schlesinger, *Developmental robotics: from babies to robots*, Intelligent robotics and autonomous agents, The MIT Press, Cambridge, Massachusetts, 2015.

[18] Angelo Cangelosi and Matthew Schlesinger, 'First Words', in *Developmental robotics: from babies to robots*, Intelligent robotics and autonomous agents, 229–274, The MIT Press, Cambridge, Massachusetts, (2015).

[19] Jean Carletta, Stephen Isard, Anne H Anderson, Gwyneth Doherty-Sneddon, Amy Isard, and Jacqueline C Kowtko, 'The Reliability of a Dialogue Structure Coding Scheme', *Computational Linguistics*, **23**(1), (1997).

[20] Rudolf Carnap, *Meaning and Necessity. A Study in Semantics and Modal Logic*, University of Chicago Press, 2 edn., 1947.

[21] *Embodied conversational agents*, eds., Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill, MIT Press, Cambridge, Mass., 2000. OCLC: 247058409.

[22] C. S. Chihara and J. A. Fodor, 'Operationalism and Ordinary Language: A Critique of Wittgenstein', *American Philosophical Quarterly*, **2**(4), 281–295, (1965).

[23] Noam Chomsky, *Syntactic structures*, Martino, Mansfield Centre, Conn., 1957. OCLC: 934673149.

[24] Leon Ciechanowski, Aleksandra Przegalinska, Mikolaj Magnuski, and Peter Gloor, 'In the shades of the uncanny valley: An experimental study of human–chatbot interaction', *Future Generation Computer Systems*, **92**, 539–548, (March 2019).

[25] Paul R. Cohen, 'If not Turing's test, then what?', *AI magazine*, **26**(4), 61, (2005).

[26] L Drescher, 'A Mechanism for Early Piagetian Learning', in *Proceedings of the AAAI*, (1987).

[27] Umberto Eco, *A theory of semiotics*, Indiana University Press, 1978. OCLC: 435571514.

[28] Robert S Epstein, Gary Roberts, and Grace Beber, *Parsing the Turing test: philosophical and methodological issues in the quest for the thinking computer*, Springer, Dordrecht; London, 2009.

[29] Richard Evans and Thomas Barnet-Lamb. Social Activities: Implementing Wittgenstein, 2002.

[30] Luciano Floridi, *The philosophy of information*, Oxford Univ. Press, Oxford, 2011. OCLC: 734050614.

[31] James Gleick, *The information: a history, a theory, a flood*, Vintage Books, New York, 1st vintage books ed., 2012 edn., 2011.

[32] Erving Goffman, *The presentation of self in everyday life*, Penguin, London, repr edn., 1990. OCLC: 832784223.

[33] H. Paul Grice, 'Logic and Conversation', in *The semantics-pragmatics boundary in philosophy*, ed., Maite Ezcurdia, 47–59, Broadview Press, Peterborough, Ontario, (2013). OCLC: 854681776.

[34] Barbara J Grosz and Candace L. Sidner, 'Attention, intentions and the structure of discourse', *Computational Linguistics*, **12**(3), 30, (1986).

[35] Stevan Harnad, 'The Symbol Grounding Problem', *Physica*, **D**(42), 335–346, (1990).

[36] Stevan Harnad, 'The Turing Test is not a trick: Turing indistinguishability is a scientific criterion', *ACM SIGART Bulletin*, **3**(4), 9–10, (October 1992).

[37] Patrick Hayes and Kenneth Ford, 'Turing test considered harmful', in *IJCAI (1)*, pp. 972–977, (1995).

[38] José Hernández-Orallo, 'AI Evaluation: past, present and future', *arXiv preprint arXiv:1408.6908*, (2014).

[39] *Believable bots: can computers play like people?*, ed., Philip F. Hingston, Springer, Berlin ; New York, 2012.

[40] Dan Jurafsky and James H. Martin, 'Dialogue and Conversational Agents', in *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*, Prentice Hall series in artificial intelligence, 829–872, Prentice Hall, Pearson Education Internat, Upper Saddle River, NJ, 2. ed., pearson internat. ed edn., (2009). OCLC: 263455133.

[41] Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra, 'Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog', in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2962–2967, Copenhagen, Denmark, (2017). Association for Computational Linguistics.

[42] Ray Kurzweil, *The singularity is near: when humans transcend biology*, Duckworth, London, 2009. OCLC: 845813950.

[43] Ian Lewin and Mill Lane, 'A formal model of Conversational Game Theory', in *In Proc. Gotalog-00, 4 th Workshop on the Semantics and Pragmatics of Dialogue, Gothenburg*, (2000).

[44] Mike Lewis, Denis Yarats, Yann N. Dauphin, Devi Parikh, and Dhruv Batra, 'Deal or No Deal? End-to-End Learning for Negotiation Dialogues', *arXiv:1706.05125 [cs]*, (June 2017). arXiv: 1706.05125.

[45] Niklas Luhmann, *Social systems*, Writing science, Stanford University Press, Stanford, Calif, 1996.

[46] David Marr, *Vision: a computational investigation into the human representation and processing of visual information*, Freeman, New York, 14. print edn., 2000. original-date: 1982.

[47] Dominic W. Massaro, Michael M. Cohen, Sharon Daniel, and Ronald A. Cole, 'Developing and Evaluating conversational Agents', in *Human Performance and Ergonomics*, 173–194, Elsevier, (1999).

[48] Michael Mateas, 'Expressive AI - A hybrid art and science practice', *Leonardo: Journal of the International Society for Arts, Sciences, and Technology*, **34**(2), 147–153, (2001).

[49] Humberto R. Maturana and Francisco J. Varela, *Autopoiesis and cognition: the realization of the living*, number v. 42 in Boston studies in the philosophy of science, D. Reidel Pub. Co, Dordrecht, Holland; Boston, 1980.

[50] John McCarthy, 'What is Artificial Intelligence?', Technical report, Computer Science Department, Stanford University, (November 2007).

[51] John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude Elwood Shannon, 'A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence', *AI Magazine (2006)*, **27**(4), 3, (December 2006). original-date: August 31, 1955.

[52] Ron McClamrock, 'Marr's three levels: A re-evaluation', *Minds and Machines*, **1**(2), 185–196, (May 1991).

[53] Colin McGinn, 'Thought and Language', in *The character of mind: an introduction to the philosophy of mind*, 83–106, Oxford University Press, Oxford ; New York, 2nd ed. edn., (1996).

[54] Marvin L. Minsky, 'A Framework for representing knowledge', Technical Report Memo 306, MIT AI Laboratory, (June 1974).

[55] Marvin L. Minsky, *Semantic information processing*, The MIT Press, Cambridge; London, 2015. OCLC: 909995563.

[56] Charles W. Morris, *Writings on the General Theory of Signs*, De Gruyter Mouton, 1971.

[57] Gina Neff, 'Talking to Bots: Symbiotic Agency and the Case of Tay', *International Journal of Communication*, **10**, 4915–4931, (2016).

[58] Barbara A. Pan and Catherine E. Snow, 'The development of conversational and discourse skills', in *The development of language*, ed., Martyn Barrett, Studies in developmental psychology, 229–249, Psychology Press, Hove, East Sussex, repr edn., (1999). OCLC: 837801515.

[59] Pandorabots, Inc. Pandorabots, 2019.

[60] Howard H. Pattee, 'Simulations, Realizations, and Theories of Life.', *Unpublished*, (1987).

[61] Jean Piaget, *The Origins of Intelligence in Children*, International Universities Press, 1952.

[62] Richard Power, 'The organisation of purposeful dialogues', *Linguistics*, **17**(1-2), (1979).

[63] David M. W. Powers, 'The total Turing test and the Loebner prize', in *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, pp. 279–280, (1998).

[64] Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. Better Language Models and Their Implications, February 2019.

[65] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, 'Language Models are Unsupervised Multitask Learners', Technical report, Open AI, (2019).

[66] Stuart J. Russell and Peter Norvig, *Artificial intelligence: a modern approach*, Prentice Hall series in artificial intelligence, Prentice Hall, Upper Saddle River, 3rd ed edn., 2010.

[67] David Schlangen, 'What we can learn from Dialogue Systems that don't work', in *Proceedings of DiaHolmia, the 13th International Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL 2009)*, pp. 51–58, (2009).

[68] Jürgen Schmidhuber, 'Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010)', *IEEE Transactions on Autonomous Mental Development*, **2**(3), 230–247, (September 2010).

[69] Friedemann Schulz von Thun, *Miteinander reden: Störungen und Klärungen: Psychologie der zwischenmenschlichen Kommunikation*, Rororo Sachbuch, Rowohlt, Reinbek bei Hamburg, originalausg edn., 1981.

[70] John R. Searle, 'Minds, brains, and programs', *Behavioral and Brain Sciences*, **3**(3), 417–457, (1980).

[71] John R. Searle, *Speech acts: an essay in the philosophy of language*, Univ. Press, Cambridge, 34th. print edn., 2011. original-date: 1969.

[72] C. Shannon, 'The bandwagon', *IRE Transactions on Information Theory*, **2**(1), 3–3, (March 1956).

[73] Claude Elwood Shannon, 'A mathematical theory of communication', *The Bell System Technical Journal*, **27**, 379–423, 623–656, (1948).

[74] Claude Elwood Shannon and Warren Weaver, *The mathematical theory of communication*, Univ. of Illinois Press, Urbana, 1998. original-date: 1949.

[75] David Harris Smith and Frauke Zeller, 'The Death and Lives of hitchBOT: The Design and Implementation of a Hitchhiking Robot', *Leonardo*, (October 2016).

[76] Luc Steels, 'The symbol grounding problem has been solved. so what's

next', *Symbols and embodiment: Debates on meaning and cognition*, 223–244, (2008).
[77] *The artificial life route to artificial intelligence: building embodied, situated agents*, eds., Luc Steels and Rodney Allen Brooks, L. Erlbaum Associates, Hillsdale, N.J, 1995.
[78] Michael Straeubig, 'On the distinction between distinction and division', *Technoetic Arts*, **13**(3), 245–251, (December 2015).
[79] Michael Straeubig, 'Let the Machines out. Towards Hybrid Social Systems.', in *Proceedings of AISB Annual Convention 2017*, pp. 28–31, Bath, (April 2017). AISB.
[80] Ilya Sutskever, James Martens, and Geoffrey Hinton, 'Generating Text with Recurrent Neural Networks', in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 1017–1024, USA, (2011). Omnipress. event-place: Bellevue, Washington, USA.
[81] Alfred Tarski, 'The semantic conception of truth and the foundation of semantics', *Philosophy and Phenomenological Research*, **4**, (1944).
[82] Tim Thornton, *Wittgenstein on language and thought: the philosophy of content*, Edinburgh University Press, Edinburgh, 1998.
[83] Michael Tomasello, *Constructing a language: a usage-based theory of language acquisition*, Harvard Univ. Press, Cambridge, Mass., 1. harvard univ. press paperback ed edn., 2005. OCLC: 254708552.
[84] Alan Turing, 'Computing machinery and intelligence', *Mind*, 433–460, (1950).
[85] Richard S. Wallace, *The Elements of AIML Style*, ALICE A.I. Foundation, Inc., March 2003.
[86] Paul Watzlawick, Janet Beavin Bavelas, and Don D. Jackson, *Pragmatics of human communication: a study of interactional patterns, pathologies, and paradoxes*, W. W. Norton & Company, New York, first published as a norton paperback 2011, reissued 2014 edn., 2014.
[87] Joseph Weizenbaum, 'ELIZA-a computer program for the study of natural language communication between man and machine', *Communications of the ACM*, **9**(1), 36–45, (January 1966).
[88] Tsung-Hsien Wen, Milica Gasic, Dongho Kim, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young, 'Stochastic Language Generation in Dialogue using Recurrent Neural Networks with Convolutional Sentence Reranking', in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 275–284, Prague, Czech Republic, (2015). Association for Computational Linguistics.
[89] Ludwig Wittgenstein, *Philosophical investigations.*, Basil Blackwell, Oxford, 2 edn., 1958.
[90] Ludwig Wittgenstein, *Tractatus logico-philosophicus*, Cosimo Classics, New York, NY, 2007. original-date: 1922.
[91] Steve Worswick. Mitsuku, 2019.
[92] Xianchao Wu, Ander Martinez, and Momo Klyen, 'Dialog Generation Using Multi-Turn Reasoning Neural Networks', in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2049–2059, New Orleans, Louisiana, (2018). Association for Computational Linguistics.