Formal Systems of Neurons in Artificial Intelligence

Molly Graham

There is growing demand for social robots to interact with humans in a variety of settings and environments, requiring these agents to possess a set of abilities to communicate and interact in a familiar manner. Included in these abilities is empathy, where certain interactions or situations will require empathic behaviours, for example, in healthcare settings. As *artificial intelligence* (AI) continues to develop, its accomplishments suggest social robots are on track to achieve certain social requirements, especially given our human tendency to anthropomorphize objects like robots and chatbots. Furthermore, the humanoid robot iCub is capable of learning to recognize objects and human emotions, suggesting it may one day express behaviours perceived as a demonstration of empathy. This notion will be discussed in detail in Chapter 3.2 of my thesis.

Here, I want to argue that robot empathy, demonstrated by iCub for example, invovles a simulation and is not a true act of empathy. This is because the robot's nervous system is generated by computer code, and as such, generates a *model* of a physical nervous system. The abstraction generated cannot fully encapsulate biological processes for the same reasons mathematician Kurt Gödel identifies in his Incompleteness Theorem: *there will exist propositions which are true and unprovable by axioms.* This connection between Gödel's Theorem and biology was developed by theoretical biologist Robert Rosen, and here I explain this abstraction in further detail. I will demonstrate that meaning and semantics cannot be fully expressed by entailment structures or syntax. This is because biological functions responsible for producing semantic meaning cannot be fully recreated in *formal systems* like computer code. This results in limitations on the kinds of behaviours expressed by various robots or AI systems using computer code, including iCub. While some behaviours may be simulable, others cannot be replicated in computerized agents given their inability to interpret the meanings humans use to communicate and socialize. Empathy requires one to adopt the perspective of another, and computerized robots cannot accomplish this because of their lack of semantic understanding. There is no way to ascertain what a human could be experiencing because the robot does not have access to semantic information used by humans. Although iCub may be able to express a simulacra of emotions like sadness, the formal models it uses cannot fully represent semantic information. Here, I will explain Rosen's distinction between *natural systems* and *formal systems* to illustrate the physical limitations of computerized robots like iCub to establish why it is not capable of genuine empathy.

Originally from Brooklyn, New York,[1] Robert Rosen received his PhD in Mathematical Biology from the University of Chicago in 1959 after studying under Nicolas Rashevsky.[2] In the early 1930's, Rashevsky had developed the first mathematical theory of neural activity generated from excitatory and inhibitory signals and all-or-nothing firing patterns.[3] Rashevsky would go on to create the journal *Bulletin of Mathematical Biophysics*[4] which would publish the notable 1943 paper by McCulloch and Pitts on the mathematical neuron.[5] While Rashevsky's approach used differential equations to model neural activity, McCulloch and Pitts used logical calculus, which Rashevsky would later admit to be a better approach.[6] By the mid 1950's, Rashevsky had taken an interest in a new approach which he termed *relational biology*, a topic which appealed to Rosen as well.[7] Biologists at the time were primarily interested in specific processes like "blood flow in arteries" and the "propagation of action potentials," however, Rosen and Rashevsky were interested in a new direction, one which studies life itself as a general phenomenon.[8] Rather than studying specific physical or structural details, relational biology instead investigates the functional and organizational features of living systems, representing them mathematically.[9] Rosen would go on to become Professor of Biophysics at Dalhousie University in 1975,[10] where he worked until he died in 1998.[11]

To understand the physical distinction between behaviours which arise from computers versus those which arise from biological systems, we must acquaint ourselves with the distinction between *natural systems* and *formal systems*. A natural system consists of some aspect of the external world or physical environment, inspiring humans to inquire about causal relationships and develop methods for scientific study.[12] Natural systems also include technologies and other man-made constructs; Rosen provides examples including "automobiles, factories, cities and the like."[13] On the other hand, a *formal* system represents or models a phenomenon generated by some natural system, as perceived by our senses or measurements, and expressed in mathematical terms. The formalism which results expresses relations between measurable properties of a natural system, where regularities or "natural laws" are generated from inductive reasoning, a generalization from a sample of occurrences.[14] While the natural

1 Rosen, 'Autobiographical Reminiscences of Robert Rosen', 2.
2 Rosen, 6–7.
3 Abraham, '(Physio)Logical Circuits', 13; Rosen, *Essays on Life Itself*, 120.
4 Abraham, '(Physio)Logical Circuits', 16.
5 McCulloch and Pitts, 'A Logical Calculus of the Ideas Immanent in Nervous Activity', 22.
6 Abraham, '(Physio)Logical Circuits', 21. See footnote 24.
7 Rosen, 'Autobiographical Reminiscences of Robert Rosen', 6.
8 Rosen, 6.
9 Rosen, *Anticipatory Systems*, 4; Rosen, *Essays on Life Itself*, 226.
10 Rosen, 'Autobiographical Reminiscences of Robert Rosen', 16.
11 Rosen, *Essays on Life Itself*, v. Judith's dedication to her father.
12 Rosen, *Anticipatory Systems*, 45.
13 Rosen, 45.
14 Rosen, 47.

world is comprised of *cause* and *effect*, a formalism describes causal relationships in terms of *entailment* or inference.[15] Rosen describes formal systems as "purely syntactic structures" which do not make use of semantic information,[16] and can thus be depicted by a variety of formats, from mathematics[17] to predicate logic[18] or input-output functions.[19] Thus, a *formalization* creates a reductive model of some aspect of a natural system, like a nervous system, to produce a model of its functionality as entailment relations,[20] seen in the McCulloch Pitts neuron.[21]

Additionally, a particular formalism cannot speak to other aspects or phenomena of the natural system it represents.[22] Elements which are beyond the scope of the model depicting other natural phenomena cannot be uncovered through the investigation of formal systems alone. This is due to their existence as *abstractions* from observation, where other phenomena arising from the same natural system must be discovered by investigating the natural system itself. Thus, formal systems are *reductive* as they view the world as an idealistic model, insofar as other variables or elements of the system are held constant, potentially ignoring important variables in the causal factors of the phenomenon in question.[23]

Between formal and natural systems resides a *modelling relation* which describes the specific relationship between a natural phenomenon and its mathematical representation.[24] The linking of properties occurs where symbols in formal systems express propositions which are true of a natural system, creating names or labels for entities or concepts identified in the external environment.[25] This *encoding* process generates a mapping or correspondence as a mathematical object, where inferential structures in formal systems must directly fit the causal relationships identified in natural systems.[26] Moving in the opposite direction, *decoding* generates predictions or hypotheses about natural systems which must be verified as true through measurement or observation. Modelling relations are established when a formalism generates accurate predictions from the theorems or axioms it employs.[27] Successful predictions thus indicate that the formalism correctly describes the natural phenomenon it is aiming to model.

---

15 Rosen, *Life Itself*, 191.
16 Rosen, 190.
17 Rosen, *Anticipatory Systems*, 25.
18 Rosen, 93.
19 Rosen, 23.
20 Rosen, 79.
21 McCulloch and Pitts, 'A Logical Calculus of the Ideas Immanent in Nervous Activity',120.
22 Rosen, *Anticipatory Systems*, 56.
23 Rosen, 262.
24 Rosen, 54.
25 Rosen, 55.
26 Rosen, *Life Itself*, 60–61.
27 Rosen, *Anticipatory Systems*, 72.

As an example of a modelling relation, Rosen produces a mathematical expression of a McCulloch-Pitts neuron on page 187 in *Anticipatory Systems.* I have included a recreation I made using LaTeX as it may further elucidate the difference between living neurons and mathematical neurons:

$$\mathrm{s(t)} = +1 \quad \text{iff} \left[\sum_{k=1}^{m} e_k(t-1) + \sum_{k=1}^{n} i_k(t-1) \right] \geq \theta;$$

$$= 0 \qquad \text{otherwise.}$$

Here, *e* represents the excitatory input variable, *i* the *inhibitory* input variable, θ is the threshold for firing, and *s(t)* represents the neuron's state at time *t* and connected to the state at the preceding instant. The neuron itself has an active state, +1 or 0, denoting the active or inactive state of the neuron.

According to Rosen, since formal systems consist of models of natural systems, the resulting behaviour is thus a simulation of the phenomenon identified in the physical environment.[28] A formalism can be simulated if its inferential structure can be expressed as a program or algorithm to be executed by a machine.[29] This indicates that a computerized robot merely simulates human behaviour by using a formalization of a neural network. Moreover, this simulation will always be a reductive, idealized model and as such, will never fully encapsulate human physiology and behaviour in computer code. Since computer code is comprised of syntactical, inferential structures, computerized robots are incapable of fully simulating human behaviour, given its reliance on semantic information. To demonstrate the gap between semantic information and syntactic structures, Rosen appeals to Kurt Gödel's *Incompleteness Theorem* to explain the inherent limitations of formalisms as entailment structures.

Gödel developed his theorem in response to David Hilbert's endeavour to reduce all mathematical truths to formalisms. In mathematics, axioms establish a foundation from which theorems can be built upon according to the rules of logic,[30] where these axioms or postulates are generally considered or assumed to be true.[31] A mathematical proof does not aim to demonstrate the truth of an axiom but instead, demonstrates how certain conclusions must necessarily follow given the rules of math and logic.[32] Thus, the question can be raised: does a mathematical system built from axioms

28 Rosen, *Essays on Life Itself,* 324.
29 Rosen, *Life Itself,* 193.
30 Nagel and Newman, *Gödel's Proof,* 4–5.
31 Nagel and Newman, 14.
32 Nagel and Newman, 12.

generate theorems which are internally inconsistent?[33] In other words, does a logical contradiction appear, within some mathematical system, from the use of various mathematical axioms? This question was the motivation for Hilbert's project, which aimed to show that the theorems of mathematical systems are indeed consistent.[34] This notion was furthered from the publication of Russell and Whitehead's *Principia Mathematica* which demonstrates how ideas within mathematics can be articulated by terms used in arithmetic.[35] It was assumed that math could be considered as a branch of logic, where all arithmetical concepts can be defined by logical truths.[36] Should this occur within a particular mathematical system, where axioms can be defined by logical truths, the system is said to be "complete." If it was not, it was believe that it could be made complete by adding more axioms to the initial list.[37] Gödel's proof would demonstrate how this is not the case, and that all mathematical systems are either incomplete and consistent, or is complete and inconsistent.[38]

To demonstrate this, Gödel devised a method for representing elements of math, such as signs, formulas, and proofs, as a unique number by transforming these elements into numerical values.[39] This unique number, called a Gödel number, acts as a specific label for a particular element; for example, the unary operator '~' which represents 'not' is Gödel number 1, while '=' which represents 'equals' is Gödel number 5. By following Gödel's rules for prescribing these unique numbers, all expressions within mathematics can be assigned a Gödel number. Problems arise, however, from self-reference. The statement "the formula with Gödel number *z* is not demonstrable" can be written as a mathematical formula, and in this case, we will label it G.[40] When the variable *z* refers to a Gödel number representing a *substitution function*, the Gödel number for G, labelled *n*, can be put into the formula. Because G represents a meta-mathematical statement, a statement about mathematics expressed in mathematical terms, it can be re-written as "the formula G is not demonstrable."[41] In this case, if G were demonstrable, then it's negation ~G would be demonstrable as well, indicating that G itself is only demonstrable if its negation is also demonstrable, generating a contradiction.[42] Since we can show that G is true through meta-mathematical means, it implies that G is true through mathematics itself, given the mapping or correspondence created between mathematics and mathematical language describing

33 Nagel and Newman, 14.
34 Nagel and Newman, 21.
35 Nagel and Newman, 42.
36 Nagel and Newman, 42.
37 Nagel and Newman, 56.
38 Nagel and Newman, 59.
39 Nagel and Newman, 69.
40 Nagel and Newman, 89.
41 Nagel and Newman, 90.
42 Nagel and Newman, 90–91.

mathematics.[43] It is considered "incomplete" because we cannot formally deduce G through mathematics itself, which leads us to conclude "if arithmetic is consistent, it is not complete."[44] In other words, there exists an infinite number of true statements that cannot be formally deduced from a set of axioms and rules of inference.[45]

Rosen appeals to Gödel's theorem to demonstrate how the relationship between semantics and syntax is analogous to the relationship between mathematics and meta-mathematics.[46] Natural language consists of both semantics and syntax because it is capable of referring to things outside of language itself, expressing ideas or concepts about these external referents and their meanings.[47] While semantics pertains to reference and meaning, syntax provides rules for altering or transforming the symbols and expressions which constitute natural language.[48] Moreover, syntax is considered to be concrete or objective, while semantics is dependent on subjective features, and as such, meanings may differ from person to person.[49] It was believed that syntax could be completely capture semantics, however, Rosen appeals to Gödel's theorem to indicate why this cannot be the case.[50] The syntax which aims to represent semantic information is incomplete, as formal systems cannot fully encapsulate natural systems. When cause-and-effect is described in terms of entailment, aspects of the natural system are removed in an attempt to identify the specific mechanics of the natural system which are responsible for a particular effect.[51] The model which results is an idealization which reduces the larger, more complex natural system into a simplified explanation of the physical world.[52] Thus, as a type of formalization, it does not *refer* to the things in the world it aims to represent, and instead, merely provides a structure or model of one particular phenomenon observed in a natural system.

Thus, semantics and meaning is *unfractionabl*e, where its functionality, in this case the referent it points out, is inseparable from the physical structures which give rise to this functionality.[53] Alternatively, something which is *fractionable* contains functionality which can be separated from its physical manifestation. The example Rosen provides is flight, where the ability to fly does not depend on the materials or shapes observed in animal wings.[54] A reference, however, cannot be fully isolated from the physical structures which support or produce it. To Rosen, the primary source of semantics is

43 Nagel and Newman, 92–93.
44 Nagel and Newman, 95.
45 Nagel and Newman, *Gödel's Proof*, 98.
46 Rosen, *Anticipatory Systems*, 69.
47 Rosen, *Essays on Life Itself*, 156.
48 Rosen, *Life Itself*, 43.
49 Rosen, *Essays on Life Itself*, 156.
50 Rosen, 157; Rosen, *Life Itself*, 44.
51 Rosen, *Anticipatory Systems*, 262.
52 Rosen, *Essays on Life Itself*, 139.
53 Rosen, 290.
54 Rosen, 291.

the external world as it is experienced or observed by individuals,[55] and as such, meaning cannot be reconstructed in purely syntactical terms.[56] Therefore, machines are incapable of processing semantic information, not only because they were not built to, but because semantic information cannot be completely represented through syntactical or entailment structures.

Using Rosen's appeal to Gödel's theorem and subsequent discussion of computer software as a simulation, it becomes apparent that the aims of developmental robotics cannot be realized. Syntax will never fully encompasses semantic information and thus, the capabilities of social robots will be insufficient for the kinds of tasks these robots are hoped to one day perform. Not only will computerized robots present a *simulation* of human communicative abilities, the simulation which does develop is not an adequate model of human behaviour. Developmental robotics makes the faulty assumption that *association* can fully replace semantics. While association is required in establishing semantic information, the way this information is presented to a biological organism is ontologically distinct from computerized robots. As Gödel demonstrates, there no amount of supplementary information, in this case computer code, which can ever fully encapsulate the semantic information relied upon by organisms.[57] As such, this semantic information does not *mean* anything to the robot; the formal system which comprises its nervous system does not, and cannot, represent the meaning of some body of information. Rosen stresses the uniqueness of biological organisms in the physical universe as a degree of physical-functional complexity unseen in other domains of natural sciences. As *anticipatory systems*, living organisms, from enzymes[58] and trees[59] to mammals like humans, generate predictive models of future events and use these predictions to alter or influence their behaviours. The study of physical processes in general, on the other hand, strictly deny that future events can influence current events,[60] indicating the uniqueness of biological organisms. In addition to these physical differences, the formal model of a living being ceases to be anticipatory and instead becomes a *reactive system*.[61] While a reactive system can potentially adjust itself based on feedback or error signals, these alterations must occur as a response to events which have already transpired.[62] Together, these reasons indicate that current approaches to AI require an understanding of physical limitations, and should a desire to overcome these limitations arise, a new paradigm for building artificial agents must be developed.

---

55 Rosen, 159.
56 Rosen, *Life Itself*, 247–48; Rosen, *Essays on Life Itself*, 292.
57 Rosen, *Essays on Life Itself*, 324–25.
58 Rosen, *Anticipatory Systems*, 320.
59 Rosen, 7.
60 Rosen, 318.
61 Rosen, 10.
62 Rosen, 40.

Returning to the themes of this project to conclude this section, it is clear that iCub is not capable of demonstrating empathy, as this ability requires an understanding of semantics to act appropriately. Without this, the act itself is not only a simulation, but a deception. Genuine empathy is not possible, and as such, we as individuals and societies must adjust our expectations for computerized AI accordingly. Although some may not be concerned or bothered by a fictitious act of empathy, others may feel betrayed or tricked especially if individuals develop feelings of attachment. We have already seen the capacity for simple AIs to trick humans into beliefs that are unfounded, as observed in the case of ELIZA, and as such, we ought to be careful about how we interpret the behaviours exhibited by various types of AI. It is possible that computerized AI will become sophisticated enough to *appear* as if a true act of empathy has occurred, however, given our analysis of empathy and Rosen's appeal to Gödel regarding semantics, it must be emphasized that this is not the case. Otherwise, individuals and societies may cultivate certain notions which are unfounded and fallacious, opening an avenue for potential harm to arise.

So despite our attempts to model biological processes to address problems in AI, as seen in the growth of studies surrounding neural networks and developmental robotics, there remains an important physical distinction between current approaches to AI and the biological organisms we aim to replicate. I suggest, based on the work of Pentti Haikonen, that artificial intelligence requires a degree of motivation for self-preservation, as this will provide an analogue of affect and thus semantic information. As such, a rudimentary version of awareness arises from associations between "sensations" and their affects on the agent's functioning. That said, any "intelligence" exhibited by this new robot remains 'artificial' because the robot is an artifact, an object of human creation, and not an anticipatory system. It is a model of an anticipatory system, a reconstruction, but remains *artificial* because it has been built by a human for a specific purpose. A brief philosophical discussion of 'artificial' will be released shortly, as it will help to motivate this argument to some degree.

★★★

Thus, as you can see, there is still a lot of work to be done to get this idea up and running smoothly. It also likely requires more explanation of Gödel's Incompleteness Theorem, which will take me some amount of dedicated time and effort in order to explicate sufficiently. There is something about Rosen's work that I find very interesting, and I would be interested to see what this can do for our understanding and conceptualization of artificial intelligence. We will be in need of good, empirical reasons for the ontological differences between AI behaviours and those exhibited by humans and animals.

Bibliography

Abraham, Tara H. ‘(Physio)Logical Circuits: The Intellectual Origins of the McCulloch–Pitts Neural Networks’. *Journal of the History of the Behavioral Sciences* 38, no. 1 (2002): 3–25. https://doi.org/10.1002/jhbs.1094.

McCulloch, Warren S., and Walter Pitts. ‘A Logical Calculus of the Ideas Immanent in Nervous Activity’. *The Bulletin of Mathematical Biophysics* 5, no. 4 (December 1, 1943): 115–33. https://doi.org/10.1007/BF02478259.

Nagel, Ernest, and James R. Newman. *Gödel’s Proof*. 2nd ed. London: Routledge, 2004. https://doi.org/10.4324/9780203406618.

Rosen, Robert. *Anticipatory Systems: Philosophical, Mathematical, and Methodological Foundations*. 2nd ed. IFSR International Series on Systems Science and Engineering, 1. New York: Springer, 2012.

———. ‘Autobiographical Reminiscences of Robert Rosen’. *Axiomathes* 16, no. 1 (March 1, 2006): 1–23. https://doi.org/10.1007/s10516-006-0001-6.

———. *Essays on Life Itself*. Complexity in Ecological Systems Series. New York: Columbia University Press, 2000.

———. *Life Itself: A Comprehensive Inquiry into the Nature, Origin, and Fabrication of Life*. New York: Columbia University Press, 1991.