Von Neumann Minds: A Toy Model of Meaning in a Natural World

Jochen Szangolies*

1 Introduction

Physical law and the teleonomy of living beings are difficult to reconcile: a stone rolls downhill because of the force of gravity, not because it wants to reach the bottom. Mental content exhibits the curious property of intentionality—of being directed at, concerned with, or simply about entities external to itself. Such 'aboutness' is arguably a prerequisite to goal-directed behavior; yet, following Brentano (1874), it is often alleged that no physical system exhibits this kind of other-directedness.

A mark on a paper is just a mark on a paper—it is only after it is interpreted as a sign, or symbol, by an intentional mind that it comes to refer to something else. Without such interpretation, there is no reference.

But interpretation itself seems to rest on the notion of reference: when we interpret the word 'apple' to refer to an apple, a reasonable suggestion seems to be that the word causes an appropriate mental representation to be called up—that is, a certain kind of mental symbol that *refers to* said apple. We are caught in a double bind: we cannot explain reference without interpretation, and we cannot explain interpretation without reference.

How, then, to reconcile the ubiquitous presence of intentionality and goal-directed behavior with blind physical law? Can we escape the double bind with a naturalist explanation, or do we have to go beyond physics, in whatever way?

In the following, I will present a 'toy model' that I argue to be capable of producing genuine meanings in a natural world. In order to introduce the model, I will, in the following section, elaborate on the nature of the double bind, introducing the *homunculus fallacy*. Following this, I will describe a model that confronts the homunculus problem head-on, eliminating the dichotomy between the representation and its (homuncular) 'user'. To do so, I will draw an analogy to the problem of reproduction, and adapt a solution due to von Neumann (1966) to the issue at hand.

2 The Homunculus in Locke's 'Camera Obscura'

Let us start by being exceedingly naive. We will take literally the suggestion made by Locke (1690, II xi 17), that

> the understanding is not much unlike a closet wholly shut from light, with only some little openings left, to let in external visible resemblances, or ideas of things without.

The mind, then, is like a *camera obscura*: a dark room, through which light only enters via a small aperture. This produces an image of the outside world on the wall opposing the hole. In more modern terms, one could imagine an external camera, whose image is projected onto an internal screen.

It is this internal image which we will consider as a mental representation. By means of this image, actions could be planned: if the image contains an apple on a plate, one could plan to reach for and eat

* jochen.szangolies@hhu.de

this apple. Thus, it seems that such an internal image suffices to implement goal-directed behavior, to plan intentional action in the outside world.

But of course, we have long since fallen prey to fallacy. When we consider how the internal image is *used* in order to formulate plans for actions in the world, we invariably postulate a *user*. This use requires recognizing the objects within the image, and representing them as apples and plates, rather than colorful blobs on a wall. But which faculty is to play the role of this user?

In attempting to give an account of the intentional machinery, we have implicitly relied on it being already present: some entity 'looks at' the internal screen, analyzes the picture, to formulate plans to enact. As the outside world is represented in this inner picture, so must the inner picture again be represented to the internal observer, the *homunculus*.

We thus must start out by immunizing ourselves against homunculi. To do so, we analyze the structure that underlies the notion of meaning. Meaning is a relation—but not, as one might believe at first, a two-place relation between a symbol and its referent, but a three-place relation: a symbol S means some referent R to an agent A. Thus, 'apple' (the symbol) means apple (the referent) to somebody sufficiently familiar with the English language, while it means nothing to somebody speaking only Chinese.

The source of the homunculus fallacy is glossing over whom a given symbol is supposed to have meaning to: we imagine that the internal picture is simply intrinsically meaningful, but fail to account for how this might come to be—and simply repeating this 'inner picture'-account leads to an infinite regress of internal observers.

Thus, if we intend to hold fast to a representational account of meaning, we must modify the underlying structure: replace the three-place relation by something that does not depend on external agency. Such a replacement structure was proposed by the present author in (Szangolies, 2015); in the following, I will introduce, as well as elaborate on, the model presented therein.

3 Von Neumann Replicators

We have seen that the main obstacle towards finding a representational theory of meaning is the three-partite structure of reference. In order to break free of this problematic structure, we will exhibit an analogous relation, that of construction, and show how the problem arises in this case. Afterwards, we discuss the ingenious solution of von Neumann (1966).

Construction is a three-partite relationship between the blueprint B, the object to be constructed O, and the constructor C. Just as a symbol S means R to the agent A, the blueprint B 'means' the object O to the constructor C. Due to this structural equivalence, we expect an analogue to the homunculus problem to surface in this context. It is not hard to see that this is, in fact, the case: consider the task of creating a self-reproducing automaton.

One possible solution is the doctrine of *preformationism*: the ancient notion that organisms contain tiny versions of the adult organisms within themselves—literal homunculi—which then grow to full size eventually. However, in order to enable indefinite reproduction—and to yield a truly faithful copy of the parent organism—, each of the tiny versions must, within themselves, already contain yet tinier versions, containing even smaller ones, and so on.

As with explaining meaning via reference, we seem to be stuck in an infinite regress in explaining self-reproductive capacities by means of the three-partite construction relation. Can we escape this conclusion?

Of course, we all know that self-reproduction in fact occurs in nature; moreover, nobody supposes that the capacity of living beings to self-reproduce depends on any mysterious, inexplicable powers. Thus, we may be confident that a solution exists.

A possibility is to just accept the homunculus regress head-on: there is, in fact, an infinite hierarchy of tiny versions stacked, in Russian doll-manner, within one another. Like the tower of homunculi gives meaning to an internal representation, this nesting enables infinite self-reproduction.

However, we do not typically believe that such infinite structures can exist in the real world. Consequently, solving both problems necessitates finding a real-world implementable replacement for the infinite structure.

A first attempt at a solution is the following. Suppose that a system is simply capable of scanning itself, producing a description that then enables it to construct an exact copy. This, at first blush, seems a very sensible solution enabling perfect self-reproduction.

However, this strategy runs into an immediate obstacle. One might anticipate this (and von Neumann did, see (von Neumann, 1966, p. 166)) due to the self-referential nature of the proposal: whenever a system scans itself, it will contain its own description as a proper part of itself; but this description then cannot be a description of the system anymore, since otherwise, it would necessarily contain a description of the system as containing a description of itself, and so on. Again, we seem to be faced with the problem of infinite regress.

Making this suggestion more concrete, Svozil (1993) proves the following:

Theorem (Svozil)**.** *In general, no complete intrinsic theory of a universal computable system can be obtained actively, i.e., by self-examination.*

The proof works by reduction to *Richard's paradox* (Richard, 1905). This paradox notes that certain English expressions unambiguously name real numbers between 0 and 1—e.g., 'the ratio of a circle's circumference to its diameter minus 3'. These expressions can be brought into lexicographic order, giving expressions of the form 'the nth English expression defining a real number' a unique meaning.

Now consider the expression 'the real number whose nth decimal is given by 9 minus the nth decimal of the real number defined in the nth expression'. This defines a unique real number r. However, it is clear that this number cannot be in the original enumeration: it differs from the nth number in the nth decimal. But nevertheless, it is clearly defined by a finite English expression! Thus, we arrive at a paradox—the original list cannot be possible, after all.

Svozil then maps all possible responses of an automaton to binary strings, and shows that there cannot be a string among these capable of reproducing all others (i.e. the sought-for intrinsic theory): if that were the case, one could construct a new string via a diagonalization method that the automaton can output such that it differs from all the others in the list.

We again recognize the three-partite structure discussed previously: we have a code (the English language), a meaning (real numbers), and a decoding-mechanism (the notion of numbers defined in English)—compare this to the symbol, its meaning, and the agent to whom it has this meaning, or to the blueprint, the object to be constructed, and the constructor carrying out this construction task. Consequently, all three problems are really the same issue under different guises: positing external agency to decipher a 'code' of some sort ultimately leads to infinite regress, thanks to issues of self-reference.

Furthermore, we could again 'solve' the issue with an infinite hierarchy: call a number 'defined$_0$' if it is named by an ordinary English expression, 'defined$_1$' if it is named by an expression referring to the notion of 'definition$_0$', 'defined$_2$' if it is named by an expression referring to 'defined$_1$', and so on. Then, r would not be a 'defined' number, but rather, it would be 'defined$_1$', solving the paradox at the cost of an infinite hierarchy of definition.

However, we need not appeal to infinite constructions to circumvent this paradox. Rather, as shown by von Neumann (1966), a construction is possible that splits the task into a syntactic and a semantic part—copying and interpreting the code. Since von Neumann's original solution was geared towards the problem of self-reproduction, we will first exhibit it in this arena, before examining how to apply it to the problem of meaning.

Von Neumann's solution incorporates the following elements. First, we have a constructor C, which, equipped with a blueprint B_O of some object O, constructs that object. Schematically, we may write:

$$C + B_O \rightsquigarrow O.$$

(Note that this should not be taken to imply that C and B_O are necessarily consumed in the process—we only indicate the newly produced entity on the right side.)

Furthermore, a duplicator D, which can duplicate any blueprint B_O, i.e. which performs the operation

$$D + B_O \rightsquigarrow B_O.$$

Additionally, we need a supervisor S, which activates first the duplicator, and then, the constructor, which leads to

$$S + D + C + B_O \rightsquigarrow O + B_O.$$

Reproduction then becomes possible by handing this assembly its own description, B_{SDC}:

$$S + D + C + B_{SDC} \rightsquigarrow S + D + C + B_{SDC}.$$

In self-reproduction, the blueprint is used in two different ways: first, it is merely regarded as a meaningless object, in order to be copied by the duplicator—only its syntactical properties are considered. Then, it is considered as a set of instructions—it is interpreted, that is, now its semantics (with respect to the constructor) are considered.

This avoids Svozil's theorem—and the homunculus problem—due to the fact that no intrinsic theory has to be obtained via self-inspection; rather, separating out syntactic and semantic aspects of the self-reproduction process enables the assembly $N = \{S, D, C, B_{SDC}\}$ to 'look at itself', rather than needing external agency. All of the elements of the three-place relation—the code, its meaning, and the agent deciphering this meaning—are now identified: the replicator N codes for itself, and reads its own code, to give rise to a copy of itself.

But von Neumann's construction enables more than mere replication. We may, for instance, change the original replicator N by adding an arbitrary pattern X to the blueprint, which yields

$$S + D + C + B_{SDCX} \rightsquigarrow S + D + C + X + B_{SDCX}.$$

That is, changing the 'genetic code' of N leads to changes of the phenotype in the next generation. Moreover, changes are hereditary—changing X to some X' will incorporate this change within the next generation. If such a replicator finds itself in competition for resources, it certain changes may enable it to better acquire them, introducing a fitness differential: replicators better able to utilize an environment's resources will reproduce at a higher rate. Consequently, these replicators have the potential to adapt to an environment. As we will see, this is a key component in how genuine meanings arise.

4 Evolving Meaning

In the previous section, we have seen how the modification of the structure underlying replication allows us to evade the homunculus problem. A von Neumann replicator N does not have to rely on external agency in order to produce a copy of itself. Moreover, these replicators are capable of 'mutation': that is, a replicator N can code for a different one N', and thus, give rise to a modified system in the next generation. The structure here is bipartite: N interpretes itself as coding for N'. In this section, we will see that this feature may be used to produce symbols that refer to things beyond themselves—without having to introduce external agents to 'decode' them. Coming back to Locke's *camera obscura*, we will see how to produce images of the world that are capable of looking at themselves—representations that are their own users.

Von Neumann originally framed his work within the language of cellular automata (CA). A cellular automaton is, essentially, a conceptual abstraction of a grid of small machines (the cells), able to interact locally with their immediate environment—that is, changing their own state based on the states of their neighbors. These machines can be combined to form larger ones, patterns of cells in different states, much as one may combine gears, levers and pulleys into a mechanism.

In the following, we will imagine an agent whose brain is given by a cellular automaton. Since there are computationally universal CA, and neuronal networks such as ordinary brains can be simulated on a computer, such an agent is in principle capable of everything that an organism equipped with a more traditional brain can do.

Now, an analogy to the *camera obscura* would then be that the CA brain, in response to external stimuli, shows a certain pattern—an 'image' of the outside world. With such a theory, we are again chasing homunculi: in order for meaning to emerge, we need some agent external to the pattern, interpreting it as pertaining to the outside world.

But we have since seen how to exorcise the homunculus: break up the three-partite structure of reference—create mental representations (CA patterns) that are their own homunculi, using themselves as symbols. This becomes possible thanks to von Neumann's construction.

A given pattern of CA cells constitutes the 'state of mind' of our hypothetical agent. Now, consider what happens if this state of mind contains a von Neumann replicator N: the replicator 'interprets' itself, giving rise to a new copy of itself—that is, the agent's state of mind has meaning to itself. Of course, this 'meaning' is of a rather trivial, self-referential sort: all that it means is merely itself.

The key to shake the agent's mind free from empty, self-referential navel-gazing is the design's evolvability. Assume that the agent is subject to certain environmental stimuli. These will have some influence upon its CA brain: they could, for instance, set up a certain pattern of excitations. As a result, the evolution of patterns within the CA brain will be influenced by these changes, which are, in turn, due to the environmental stimuli.

Now imagine that there is a population of replicators active within the CA. Different designs will then unavoidably perform differently well in different environments—some might loose their ability to self-replicate completely; others, in contrast, might experience a boost, thus becoming more and more frequent within the population. Moreover, changes introduced within a replicator pattern may influence their replication; that is, these patterns will evolve towards a form more suited to the current conditions.

Effectively, the outside environment determines the *fitness landscape* for replicators in the CA-brain—they dictate which replicators enjoy reproductive success, and to what degree.

Consequently, confronted with a certain environmental situation, within the agent's CA brain, a replicator population suitably adapted to the CA fitness landscape will gradually become dominant.

Two key points must then be made here: first of all, through the process of mutation, the replicator comes to mean something beyond itself—it interpretes itself as a different pattern, again in a way independent of any outside agency. Second, the evolutionary process brings the replicator into ever-closer correspondence with the outside environment: the better adapted a replicator becomes to the CA conditions set up by the environment, the more it becomes a function of these conditions—and with that, of the environment.

Selection processes leave traces on those entities subject to them that are characteristic of the environment to which they are adapted. A dolphin's streamlined body attests to its living in a fluid medium; a fish's atrophied eyes bear witness to its having moved from an environment containing light to a lightless one, say due to a population becoming isolated in a cave that closed itself off. In evolutionary processes, the environment 'informs' organisms in the sense of giving shape to them.

Taken together, these two points mean that replicators in a CA-brain that is subject to environmental influences gradually come to be *about* that environment—they interpret themselves as patterns whose form comes to be ever more adapted to this environment, thus reflecting it.

Douglas Hofstadter coined the dictum that "[t]he mind is a pattern perceived by a mind" (Hofstadter and Dennett, 1981, p. 200). In the above, we have presented a mechanism to realize this notion, in toy-model form: employing a replicating structure that interprets itself as something different from itself—that has meaning to itself, the way that the content of a mind has meaning to said mind.

Suppose now that the dominant replicator at a given time becomes capable of directing the agent's actions. It is immaterial here to speculate on how, exactly, this process might work—the only thing that matters is that the dominant replicator, best adapted to the CA-fitness landscape, and consequently, to the outside environment, is put into the driver's seat.

Due to the connection between a replicator and the environment, the 'active symbols' formed in this way can be employed just as homunculi might be employed to use the representation of the external world on an internal viewscreen. A symbol created in an environment containing, e.g., an apple on a plate may produce actions appropriate to the presence of that apple—i.e. cause the agent to reach for the apple and take a hearty bite.

We may imagine the symbol itself to be grabbing for the apple: just like an operator of a crane lends their intentionality to it, making it engage in purposeful behavior like lifting a load to the third floor, the symbol interpretes its own form as information about the environment, lending its agency to the agent it controls. In this sense, intentionality is contagious: intentional agents can lend their goal-directedness to larger systems they are part of.

In this way, the symbol becomes itself a kind of homunculus—albeit in a non-problematic way: while the agent's intentionality is derived from the symbol's, there is no infinite regress, due to the capacity of the symbol to interprete itself, and cause appropriate actions in the agent based on this interpretation.

Of course, the model as presented is highly speculative. There is, to the best of my knowledge, no experimental evidence for anything like these self-replicating configurations in real, organic brains—although theoretical proposals for self-replicating patterns of neurons do exist (Fernando, Karishma, and

Szathmáry, 2008). However, the above considerations may become more plausible upon realizing that they serve to solve further problems that otherwise seem difficult to account for.

One is the so-called *problem of error* (Dretske, 1981). This problem consists in the recognition that on many representational theories of mental content, it is difficult to account for erroneous judgments, e.g. for believing that something is present when it is in fact not. An example of this would be thinking that a strange person is present in a dark room, when in fact, there is only a jacket hung on the wall.

A theory on which a representation is supposed to mean whatever has caused its activation—that is, on which a mental symbol means 'apple' if it is triggered by the presence of apples—immediately falls prey to this issue.

If there is some mental symbol that is triggered whenever I see a strange person in a dark room—producing the appropriate response of being startled—, and it is triggered by a jacket, on such a theory the symbol does not mean 'strange person', but rather, 'strange person or jacket', and its being triggered is completely appropriate. But I don't find myself believing that there is a 'strange person or jacket' in the room, and having this belief confirmed by the presence of a jacket; rather, I find myself believing that there is a strange person in the room, a belief that will then, to my relief, be disconfirmed by finding that it is only a jacket.

The model as presented above accounts for this easily: while at first, a certain replicator causing a startling response was dominant within the mental population, it simply was not the most well-adapted to the actual environmental conditions, becoming eventually replaced by one fitting them better.

The other problem is connected with the question of why such a seemingly baroque scheme might develop within organisms at all. A possible reason for this is that these organisms face the challenge of coping with an open-ended environment—that is, with a possibly limitless set of conditions. In artificial intelligence, this gives rise to the so-called *frame problem* (McCarthy and Hayes, 1969): a robot navigating an environment needs to possess information about this environment—roughly, about what kinds of problems to expect, and how objects behave, in order to solve them. But in the natural world, the set of object-behaviors and problems an organism might be faced with is not clearly delimited, and potentially infinite.

An evolutionary approach, in contrast, is capable of adapting to arbitrary environmental conditions. Thus, organisms possessing such a mechanism have an inherent advantage over organisms lacking it, and hence, enjoy greater reproductive fitness—leading, as a final consequence, to the existence of meaning, aboutness, and goal-directed behavior in the world.

An interesting parallel may be drawn here to the immune system: the so-called *clonal selection theory* due to Burnet et al. (1959) postulates that the diversity of antibodies to combat an infection is due to a Darwinian process. In this way, the immune system does not need access to a near-limitless variety of appropriate 'responses' to all conceivable infectious agents, but rather, may evolve an appropriate response when necessary. Consequently, a strategy of using a Darwinian adaptation process in order to produce an appropriate reaction to near-limitless environmental variety already exists within nature's toolkit—and, as many examples of convergent evolution demonstrate, nature loves to recycle its solutions. (Indeed, the existence of this evolutionary immune response inspired the neuronal replicator model of Fernando, Karishma, and Szathmáry (2008), via the *neural Darwinism* of Edelman (1987).)

5 Conclusion

I have presented a toy model of how meaning, aboutness, and intentionality emerge in a natural world. The model's key insight is that, in order to solve the homunculus problem, the three-partite structure of reference—a symbol (or code) means its referent to an agent—must be broken up, since otherwise, we face an infinite regress of interpretational agencies.

I have argued that such a breaking up can be achieved by a mechanism analogous to von Neumann's self-replicating machines. Construction is likewise based on a three-partite structure—a blueprint becomes 'translated' into a certain object by a constructor. Von Neumann's design breaks up this structure, avoiding the infinite regress by creating objects capable of both copying and reading blueprints of themselves contained within themselves. Furthermore, the design enables evolvability: a von Neumann replicator N may construct a different object N' from the code within itself; N' then may itself be a replicator, possibly better adapted to its environment.

Translated back into the realm of symbols and its referents, N constitutes a symbol interpreting itself as N'—i.e. the symbol and the agent to whom it has meaning are identified, eliminating the homunculus, while nevertheless producing an interpretation of symbols.

The model then postulates that symbols come to 'embody' information about the outside world by an evolutionary process: environmental conditions, via the senses, set up a fitness landscape in an agent's 'brain', dictating which replicators—which symbols—enjoy the greatest reproductive success. In this way, the environment informs the replicators—differences in the environment lead to differences in the reproductive fitness of replicators.

Once a certain replicator has become dominant, it becomes capable of influencing the agent's actions, in such a way as to be consistent with the information about the environment that it embodies. Thus, the environment may include an element X; this causes a certain replicator with the trait x to become dominant; the replicator then interpretes itself, in turn causing actions in the agent appropriate to the presence of X. Just like a driver knows about the conditions on the road, and where to steer, the symbol knows about X, and causes the agent to act accordingly—it lends its intentionality to the agent, in a way that does not cause an infinite regress of intentional agents.

Furthermore, I have argued that the model makes headway in addressing the problem of misrepresentation and the frame problem. Misrepresentation is caused by a replicator becoming momentarily dominant that is not best-adapted to the environment, to then be replaced by a better-adapted one; furthermore, one can make a case that evolutionary approaches are capable of adapting to arbitrary environmental situations, thus alleviating the frame problem.

Despite these apparent successes, the model still leaves open problems. Some issues relate to a more precise formulation of the model itself—a precise formulation of how exactly the environment determines the fitness landscape is still outstanding. Additionally, it is not clear how the dominant replicator is selected in order to guide behavior.

Further questions relate to the implementation of the model in biological organisms. It is not clear whether organic brains actually support the replicating structures the model needs—that is, whether they are biologically capable of implementing the model in a sufficiently efficient way (that they are capable of implementing the model in principle is shown by the computational equivalence of cellular automata and neural networks).

Furthermore, there are issues regarding whether the model actually captures all of the phenomena associated with meaning. One question here is the generation of compound symbols: how do symbols for 'coffee' and 'mug' combine to form a symbol for 'coffee in the mug'? Or is there a separate symbol for such a compound entity? A promising direction here may be to again look to well-established biology as an inspiration: after all, we readily know examples of 'compound replicators'—namely, multicellular beings. Might such a strategy also work for symbol composition?

References

Brentano, Franz (1874). *Psychologie vom empirischen Standpunkte.* Berlin: Duncker & Humbblot.

Burnet, Sir Frank Macfarlane et al. (1959). *The Clonal Selection Theory of Acquired Immunity.* Nashville: Vanderbilt University Press.

Dretske, Fred (1981). *Knowledge and the Flow of Information.* Cambridge: MIT Press.

Edelman, Gerald M (1987). *Neural Darwinism: The Theory of Neuronal Group Selection.* New York: Basic Books.

Fernando, Chrisantha, KK Karishma, and Eörs Szathmáry (2008). "Copying and Evolution of Neuronal Topology". In: *PloS one* 3.11, e3775.

Hofstadter, Douglas R and Daniel C Dennett (1981). *The Mind's I: Fantasies and Reflections on Self & Soul.* New York: Bantam Books.

Locke, John (1690). *An Essay concerning Humane Understanding.* London: T. Basset, E. Mory.

McCarthy, John and Patrick J Hayes (1969). "Some Philosophical Problems from the Standpoint of Artificial Intelligence". In: *Readings in artificial intelligence*, pp. 431–450.

Richard, Jules (1905). "Les principes des mathématiques et le probleme des ensembles". In: *Revue générale des sciences pures et appliquées* 16.541, pp. 295–6.

Svozil, Karl (1993). *Randomness & Undecidability in Physics.* Singapore: World Scientific.

Szangolies, Jochen (2015). “Von Neumann Minds: Intentional Automata”. In: *Mind and Matter* 13.2, pp. 169–191.

von Neumann, John (1966). *Theory of Self-Reproducing Automata.* Ed. by Arthur W. Burke. Champaign: University of Illinois Press.