



Self-Reference, Self-Representation, and the Logic of Intentionality

Jochen Szangolies¹

Received: 7 June 2020 / Accepted: 18 September 2021 / Published online: 23 November 2021
© The Author(s), under exclusive licence to Springer Nature B.V. 2021

Abstract

Representationalist accounts of mental content face the threat of the homunculus fallacy. In collapsing the distinction between the conscious state and the conscious subject, self-representational accounts of consciousness possess the means to deal with this objection. We analyze a particular sort of self-representational theory, built on the work of John von Neumann on self-reproduction, using tools from mathematical logic. We provide an explicit theory of the emergence of referential beliefs by means of modal fixed points, grounded in intrinsic properties yielding the subjective aspects of experience. Furthermore, we study complications introduced by allowing for the modification of such symbolic content by environmental influences.

1 Introduction

A naive view of mental content is provided by a literal reading of Locke’s allegory of the mind as a *camera obscura* (Locke 1690)—a darkened room, into which light from the outside shines through a small aperture, producing an image on the opposing wall. This image could be considered a representation of the outside world: a picture of an apple on a table, say, could be used to formulate a plan to grab the apple to eat it. In this way, a certain goal—acquiring nourishment—combined with beliefs either explicit in the representation—there is an apple on the table—or implicit in the cognitive structure of the agent—apples are nourishing—is exemplified in the taking of a certain action—grabbing for the apple.

This account, though simple, will guide us in the following. But first, we need to take heed of an obvious paradox contained within: in appealing to the image on the wall being *used* to formulate a plan, we have implicitly introduced a *user*—but have not taken account of how such a user could make use of this representation. As described, the representation itself is supposed to be the vehicle by which the content of the

✉ Jochen Szangolies
jochen.szangolies@gmx.de

¹ Independent, Staffelsbergstr. 100, 50765 Cologne, Germany

external world enters the mind; how we would make use of an external representation—such as a map, or indeed an image projected onto a wall—is then by means of an internal one. But how does the implicit user of this internal representation make use of it, in turn? If we suppose that the mechanism is iterated, we have succumbed to the *homunculus fallacy*—explaining a capacity in terms of itself, the explanation is circular and void of content. Williford (2006) calls this the ‘first regress’.

One way out of this dilemma is to collapse the representation and the representation’s user into one—that is, have a conscious state be subject to its own representational capacities. This introduces, however, the ‘second regress’ (Williford 2006): a conscious state, in self-representing, would have to represent all of its representational properties—and thus, its property of representing itself, or representing its property of representing itself, and so on.

To address this sort of difficulty, in (Szangolies 2015, 2018, 2020) the notion of *von Neumann Mind* was introduced and developed. On a related note, in the setting of biosemiotics, studying the emergence of meaningful information from molecular substrates, von Neumann’s construction is also appealed to by Pattee (1969), and developed in a large body of work collected in (Pattee and Rączaszek-Leonardi 2012). Waters (2012) investigates the applicability of this framework to the study of human language. Additionally, the idea that von Neumann’s construction could have a role to play in the emergence of the mind is briefly alluded to in (Hofstadter 2007, chap. 20).

In brief, the intent is to adapt the work of von Neumann (1966) to eliminate problematic self-reference from the problem of reproduction to address the difficulties of representation. Reproduction faces its own homunculus regress: a naive theory of how an organism might self-reproduce has it containing a smaller version of itself—this is the doctrine known as *preformationism*. Its obvious defect is that either, the offspring will then be unable to reproduce itself—or else, must come equipped with tiny internal copies all its own, and so on: the ‘second regress’ looms.

Von Neumann saw a solution of this dilemma, by replacing the copy with a blueprint, and dividing the action of reproduction into a semantic and a syntactic part—the blueprint is used to construct a daughter agent, which is then fitted with a copy of it. In Sect. 2, the model will be introduced in greater detail, and the proposal for how to apply it to exorcise the homunculus from mental content presented. Furthermore, as a test case, the application to the implementation of a computation in a physical system, discussed in (Szangolies 2020), is presented in outline.

While the model as developed so far has a certain promise in eliminating the self-referential circularity of the homunculus, it has not, as of yet, been explicitly articulated how, precisely, von Neumann replicators ultimately can come to refer to, or be about, matters of fact beyond themselves. This lacuna is to be addressed here, in Sect. 3, by means of the notion of *modal fixed points*—self-referential formulas that are provably equivalent to formulas with the self-reference eliminated. Before presenting this solution, however, we must grapple with the problems introduced by Löb’s theorem (Löb 1955)—that, stated informally, no proof system can trust its proofs in general, unless an actual proof has been found.

In Sect. 4, an analogy due to Seager (2016), connecting structural or relational properties to axioms of a formal system, and intrinsic properties with its model, in the sense of mathematical logic, is presented, to argue that the access to intrinsic

properties offered by the self-referential nature of the von Neumann process yields the capacities necessary to ground the infinite regress, which thus provide the means by which reference is instantiated.

On this basis, we will study how the influence of environmental stimuli affects representations in the form of adaptations to successive generations of replicators. There, we will meet the homunculus regress in a new guise, and elucidate how the proposal of subjective experience as awareness of intrinsic properties from (Szangolies 2020) can provide a way of bottoming out, without introducing the need for an ever-increasing tower of formal systems of greater deductive power as in (Yudkowsky and Herreshoff 2013).

Finally, we review the proposal and give some outline for future development in Sect. 5.

2 Von Neumann Minds: Self-reading Symbols

We start with a brief exposition of von Neumann's design for a self-reproducing automaton (von Neumann 1966). Von Neumann's original setting was that of cellular automata (CA), but the formalism is independent of this setting. It is important, however, to note that an *automaton* in this sense can be thought of as, essentially, a concretely instantiated pattern: a pattern of cells in various states, in the CA-case, a pattern of certain physical components (think transistors, sensors, actuators, and the like), or even a pattern of electrochemical excitations in neural tissue—the latter being the kind of implementation implicitly lurking in the background of the present considerations.

The key notion is that of a *universal constructor*: a particular automaton \mathcal{U} that, if given a suitable description ('theory') $\mathcal{T}_{\mathcal{X}}$ of another automaton \mathcal{X} , acts such as to construct an instance of \mathcal{X} :

$$\mathcal{U} + \mathcal{T}_{\mathcal{X}} \rightsquigarrow \mathcal{X}$$

By convention, only newly added elements will be indicated to the right of the \rightsquigarrow -symbol (read: 'constructs'). This should not be taken to indicate that the elements to the left are necessarily consumed in the process—although they may be, for instance, in a scenario where an automaton replaces itself with a superior successor-version.

Von Neumann's conception of the universal constructor owes a debt to Turing's notion of the *universal (computing) machine* (Turing 1937), a device capable of implementing arbitrary computations upon being given a finite 'recipe' (i. e., a program) for doing so. Both then share this characteristic of carrying out arbitrary tasks within a given domain (construction vs. computation) from a finite specification of this task. However, the universal constructor differs in an important respect: its input, the device itself, and its output are all elements of the same domain—cellular automaton patterns, or physical objects—whereas the universal computer is typically considered to implement syntactical operations on vehicles carrying symbolic content. As we will see later on, this poses a problem for the question of which computation a physical system

implements, leading to worries of triviality for computational theories of the mind (Putnam 1988; Searle 1992).

Adjoined to the constructor \mathcal{U} is a *duplicator* \mathcal{D} , capable of duplicating any description:

$$\mathcal{D} + \mathcal{T}_{\mathcal{X}} \rightsquigarrow \mathcal{T}_{\mathcal{X}}$$

Finally, we introduce a *supervisor* \mathcal{S} , which governs the action of \mathcal{U} and \mathcal{D} , activating first one, then the other, such that any description $\mathcal{T}_{\mathcal{X}}$ will be both translated into an actual pattern, and copied:

$$\mathcal{U} + \mathcal{D} + \mathcal{S} + \mathcal{T}_{\mathcal{X}} \rightsquigarrow \mathcal{X} + \mathcal{T}_{\mathcal{X}}$$

If we now consider the von Neumann automaton $\mathcal{N} = \mathcal{U} + \mathcal{D} + \mathcal{S}$, and supply it with its own description $\mathcal{T}_{\mathcal{N}}$, we obtain

$$\mathcal{N} + \mathcal{T}_{\mathcal{N}} \rightsquigarrow \mathcal{N} + \mathcal{T}_{\mathcal{N}},$$

that is, the self-replication of the pattern $\mathcal{R} = \mathcal{N} + \mathcal{T}_{\mathcal{N}}$.

Via the separation of the replicative action into the steps of construction and copying—the semantic and syntactic evaluation of the description $\mathcal{T}_{\mathcal{N}}$ —, the homunculus is therefore banished from the problem of reproduction.

The proposal of (Szangolies 2015, 2018) then was to translate this solution into the mental realm—that is, to put the replicator into an agent’s brain, where it acts as a representation of the environment, and governs its behavior appropriately. This needs two further notions: the evolvability of von Neumann’s design, and the ability to self-inspect, and base actions upon the results of this self-inspection.

The first of these is achieved by noting that we can modify the description of a replicator before the reproduction step, leading to the construction of a slightly adapted replicator:

$$\mathcal{N} + \mathcal{T}_{\mathcal{N}'} \rightsquigarrow \mathcal{N}' + \mathcal{T}_{\mathcal{N}'},$$

The construction of adapted/altered versions of a replicator simply takes heed of the fact that we sometimes, certain indications to the contrary notwithstanding, do change our minds—that is, that the successor mind-state (i. e. replicator pattern) to the current one differs from it in salient ways.

We might, for example, consider the addition of an arbitrary pattern \mathcal{E} to the description in the ‘parent’ generation,

$$\mathcal{T}_{\mathcal{N}} \longrightarrow \mathcal{T}_{\mathcal{N}'} = \mathcal{T}_{\mathcal{U} + \mathcal{D} + \mathcal{S} + \mathcal{E}},$$

leading to its expression in successive generations:

$$\mathcal{N} + \mathcal{T}_{\mathcal{N}'} \rightsquigarrow \mathcal{U} + \mathcal{D} + \mathcal{S} + \mathcal{E} + \mathcal{T}_{\mathcal{U} + \mathcal{D} + \mathcal{S} + \mathcal{E}}$$

This then introduces a kind of heritable ‘mutation’ of the overall automaton. This mutation, we stipulate, contains information about the environment, say as impressed upon the agent’s brain-state via whatever sensory modalities—but note that I will take a wide view of the term ‘environment’, including, for example, such things as past experiences and other ‘external’ influences that might impinge on a mental state, including for instance affective states of an organism brought on by certain emotions. It is then this that makes the von Neumann replicator into a mental representation.

Additionally, changes to the tape $\mathcal{T}_N \rightarrow \mathcal{T}_{N'}$ might lead to the construction of a changed—and perhaps, improved in some respect—successor automaton \mathcal{R}' , with abilities exceeding that of \mathcal{R} .

However, what is needed further is the capacity to act on the basis of what is being represented—the capacity, in other words, to *use* itself as a representation. This enters into the picture due to the fact that the universal constructor, as von Neumann envisions it, likewise has access to a universal computational device. In consequence, it is able to prove arbitrary theorems, and, since it has access to its own formal description—its ‘code’, so to speak—it is able to prove arbitrary theorems about itself.

In particular, it is able to prove theorems regarding which actions will enable it to best achieve a certain goal G , given the information about its own state—and thereby, as encapsulated by the environmentally-induced pattern \mathcal{E} , its knowledge of the world outside. More explicitly, G refers to a certain state of the world, in which the aim of the agent has been achieved; its goal is then to bring about that state. The aim, then, is to show that the replicator could prove that, given \mathcal{E} , $G =$ “obtaining nutrition” may be achieved by executing a certain kind of grabbing motion, followed by a biting-into of the grabbed item, which would license the conclusion that the replicator—or more accurately, the agent whose mind-state is given by the replicator—believes that the grabbed-and-bitten item is something nutritious located at such-and-such a position in the world—say, an apple on the table.

In the general case, the carrying out of an action, we stipulate, is facilitated by the production of a certain ‘action pattern’ \mathcal{A} (which may or may not itself be a replicator, or subset of one). In the scenario where we think of the patterns as physical automata, this might be the building of a special sort of tool—perhaps, for instance, to more efficiently acquire certain raw materials needed in construction. Thinking about the patterns as representing some agent’s brain state, an action pattern might cause the agent to carry out a certain action—say, by setting up the right pattern of excitations in the motor cortex, or some equivalent. Reproduction is then a special case of action where the action pattern \mathcal{A} is the original replicator, \mathcal{R} , or some modified version \mathcal{R}' .

Making this fully explicit will, however, introduce some additional complications, which we need to develop some further methods to address.

It is important to note that the proposal as outlined here differs in some respects from the original one put forward in (Szangolies 2015, 2018). Many of these changes are superficial, having little impact on the model’s core, but some are more substantial. To better orient the reader, we will endeavour to make these differences more explicit.

The core component of the von Neumann Mind is its attempted solution of the homunculus regress plaguing representationalist accounts of intentionality by means of providing a self-representing structure collapsing the distinction between a representation and its user. This can be achieved in different ways; however, the differences

between these are ultimately of little impact on the formal structure of the model to be studied here. Von Neumann originally proposed a structure strictly separated into an active and a passive part—an automaton and the tape containing its description.

Later proposals have chosen different realizations; notably, the original implementation in (Szangolies 2015) was based on a proposal due to Laing (1977), which initially does not contain an explicit description of the entire automaton, but merely that of an analyzer, which can create descriptions of arbitrary patterns. However, this difference does not alter the overall structure of the model, as in both cases, replication takes place by separating the process into duplication (of the tape) and construction.

More substantially, the present work aims to give a novel account of how representations are connected to their representational contents. The von Neumann replicator represents, first and foremost, itself; it thus becomes a question how to ‘redirect’ the representational link to become outward-facing, to anchor representation in the outside world. Originally, this was proposed to work via a ‘meaning-as-action’-theory closely connected to the *success semantics* of Whyte (1990): the object of a representation is that towards which the actions it causes an agent to perform are directed.

But the von Neumann Mind does not entail a commitment to such a theory of meaning. Rather, we develop, in Sect. 4, a novel proposal. This will take heed of an innovation introduced into the model in (Szangolies 2020): where originally, phenomenal qualities played no role for the von Neumann Mind, there, it was proposed that the self-referential structure of the von Neumann construction could furnish access to its own intrinsic (as opposed to structural) properties. These, then, can be used to ground representations. Developing this suggestions into a full-fledged theory will be our main focus here.

2.1 Quines and Replicators

Von Neumann’s replicator design bears a close resemblance to what is called a *quine* in computing—that is, a program whose sole output consists of its own source-code. The name ‘quine’ was coined by Hofstadter (1979) in reference to Quine’s paradox (Quine 1976), which notes that the sentence

“yields falsehood when preceded by its quotation” yields falsehood when preceded by its quotation

cannot be assigned a consistent truth value.

The similarity lies in the establishment of self-reference by means of quotation, mirroring that of the Gödel sentence, where a quotation scheme in the form of a *Gödel numbering* is used to establish self-reference.

An example for a simple quine is given in Listing 1.

```

function Quine1 {
$start = '@'
$end = "'@"

# Separation character
$sep = [char]59

$tape = @'
function Quine1 {
$start = '@'
$end = "'@"

# Separation character
$sep = [char]59

$tape = ;

# Construct the first lines up to $tape = ...
$tape.Split($sep)[0] + $start | Write-Host

# Copy the tape
$tape | Write-Host

#Construct the final lines
$end + $tape.Split($sep)[1] | Write-Host
}
'@

# Construct the first lines up to $tape = ...
$tape.Split($sep)[0] + $start | Write-Host

# Copy the tape
$tape | Write-Host

#Construct the final lines
$end + $tape.Split($sep)[1] | Write-Host
}

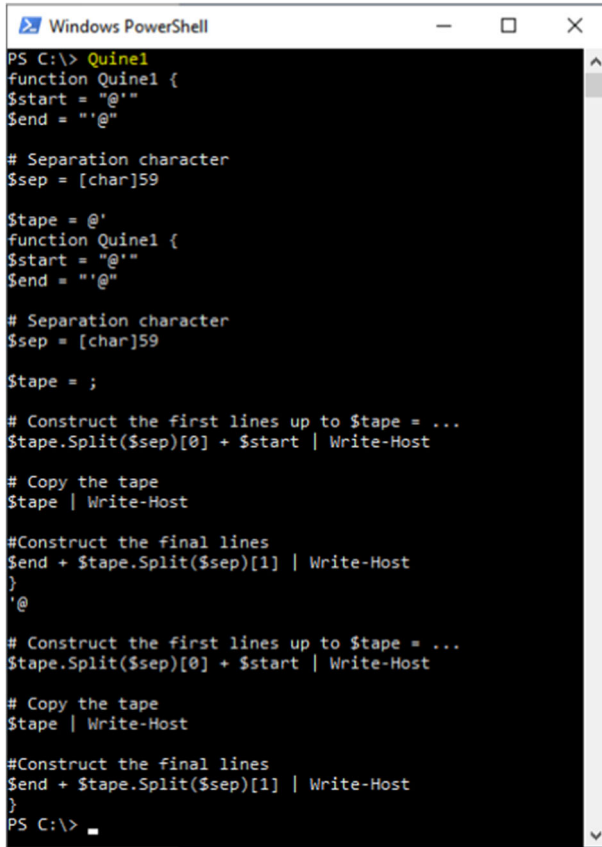
```

This is written in the scripting language PowerShell, which is present on every modern computer running the Windows operating system. The script provides a function named ‘Quine1’; calling this function will then simply output the function’s code.

The process by which this ‘self-reproduction’ is achieved closely parallels the case of the von Neumann replicator: first, the lines at the beginning of the function are read off from the ‘tape’ stored in the variable $\$tape$, up to the separator ‘;’, and the symbol @’, indicating the start of a multiline string (the tape) is inserted. Then, the tape is copied as-is, and afterwards, the string closing symbol ’@ is inserted, followed by the transcription of the rest of the lines of code. The output of the program is shown in Fig. 1.

One reason for including this example is to make vivid the point that it is not \mathcal{N} (i. e. the code of the program except for what is stored within $\$tape$) that is replicated, but the full $\mathcal{R} = \mathcal{N} + \mathcal{T}_{\mathcal{N}}$.

In the same vein, one could imagine a more elaborate example of such a program, that uses the quining method to access its own code; doing so, it can then proceed to prove arbitrary theorems about itself, if it contains the requisite theorem-proving



```

PS C:\> Quine1
function Quine1 {
$start = "@"
$end = "'@"

# Separation character
$sep = [char]59

$stape = @'
function Quine1 {
$start = "@"
$end = "'@"

# Separation character
$sep = [char]59

$stape = ;

# Construct the first lines up to $stape = ...
$stape.Split($sep)[0] + $start | Write-Host

# Copy the tape
$stape | Write-Host

#Construct the final lines
$end + $stape.Split($sep)[1] | Write-Host
}'
'@

# Construct the first lines up to $stape = ...
$stape.Split($sep)[0] + $start | Write-Host

# Copy the tape
$stape | Write-Host

#Construct the final lines
$end + $stape.Split($sep)[1] | Write-Host
}
PS C:\>

```

Fig. 1 Output produced by calling the Quine1-function

machinery. These will pertain to its own properties—and again, to properties of the complete assembly, not merely to the data present on its tape, for instance. A simple example of such code is given in Appendix A.

Using this method of ‘quining’ (to be well-distinguished from ‘quining’ as used by [Dennett 1988]), this program performs two actions on its own code—or more accurately, the code of a resulting offspring. Rather than merely printing its own code to the screen, this version stores it in the variable `$quine`, thus creating a virtual copy of itself—complete with tape—it can then examine at leisure. Then, the length of the content of the variable—and thereby, of the total program—is output to the screen.

Furthermore, the program carries out a check, to see whether its code contains the character ‘?’. If it does, it will write to the screen that a copy will be produced, and return the content of the variable `$quine`; otherwise, nothing is returned, and the program writes ‘This program will not produce a copy of itself.’ The program has thus access to its own properties, and its own behavior.

The other reason for giving this example is perhaps the more important one, which is to highlight the *difference* between such self-reproducing programs and von Neumann

replicators: while the code above relies on a suitable execution environment to produce a copy, and is, otherwise, an inert string of symbols, the von Neumann replicator is itself an active agent, acting on itself and its surroundings to produce a copy of itself. In a quine, the ‘constructor’ is essentially provided by the execution environment, while it is a proper part of a von Neumann replicator.

This also harkens back to the distinction between a universal Turing machine and von Neumann’s universal constructor, where the former acts on a domain of symbols, while itself being an (idealized) physical apparatus, while the latter is an element of the same domain it acts on.

A word of caution is in order, however. We must not overinterpret the self-proclamations of such a quine. After all, just because we can get a program to write ‘Socrates is mortal’, it does not follow that there is an agent who *believes* that Socrates is mortal anywhere to be found. The production of signs is not the production of meanings; a program with access to its own source code does not thereby acquire knowledge of its own self.

2.2 Implementing a Computation: A Test Case

After the preceding interlude, let us illustrate the model so far by means of an example taken from (Szangolies 2020): the question of which computation a certain device implements. In everyday usage, this does not seem a particularly interesting question: if I push the requisite sequence of buttons on my calculator—‘2’, ‘+’, ‘3’, ‘=’—and am presented with a certain display—‘5’—in response, I will take this as sufficient indication that the calculator has computed the sum of the numbers 2 and 3, correctly yielding 5.

But matters are not nearly always so clear-cut. In fact, a long line of so-called *triviality arguments* (Godfrey-Smith 2009) allege the conclusion that there may be no fact of the matter regarding what a given system computes, or even whether it computes at all. The two most famous examples of this sort of argument are due to Searle (1992), who argued that there is sufficient complexity in the microphysical dynamics of his office wall for it to be considered to implement the Wordstar-program, and Putnam (1988), who gave an explicit formal mapping of the sequence of states a stone in the sun traverses to those of an arbitrary finite state automaton.

In the same vein, in (Szangolies 2020), an argument is presented to the effect that any given system at most fixes the structure of the computation it implements, which, however, does not suffice to fully individuate that computation. Computation, it is argued there, is in fact an instance of modeling, and just as the same system can be a model for different objects, so too can the same system implement different computations.

Consider thus a box presenting, on its front, four switches, and three lamps. The lamps come on in a certain pattern once the switches have been set (we may consider a ‘go’ button, or a time delay, to differentiate between initial and final states of the system). We can write the total state of the box, at each instance of time, as a septuple of the form $(s_{11}, s_{12}, s_{21}, s_{22}, l_1, l_2, l_3)$, where s_{ij} refers to the state of one of the four switches, taking the values \uparrow (‘switch up’) and \downarrow (‘switch down’), while l_i denotes the

Table 1 State-transition table for the box

S_{in}	S_{fin}
(↓, ↓, ↓, ↓, ●, ●, ●)	(↓, ↓, ↓, ↓, ●, ●, ●)
(↓, ↓, ↓, ↓, ↑, ●, ●, ●)	(↓, ↓, ↓, ↑, ●, ●, ○)
(↓, ↓, ↑, ↓, ↓, ●, ●, ●)	(↓, ↓, ↑, ↓, ●, ○, ●)
(↓, ↓, ↑, ↑, ↓, ●, ●, ●)	(↓, ↓, ↑, ↑, ●, ○, ○)
(↓, ↑, ↓, ↓, ↓, ●, ●, ●)	(↓, ↑, ↓, ↓, ●, ●, ○)
(↓, ↑, ↓, ↑, ↓, ●, ●, ●)	(↓, ↑, ↓, ↑, ●, ○, ●)
(↓, ↑, ↑, ↓, ↓, ●, ●, ●)	(↓, ↑, ↑, ↓, ●, ○, ○)
(↓, ↑, ↑, ↑, ↓, ●, ●, ●)	(↓, ↑, ↑, ↑, ○, ●, ●)
(↑, ↓, ↓, ↓, ↓, ●, ●, ●)	(↑, ↓, ↓, ↓, ●, ○, ●)
(↑, ↓, ↓, ↑, ↓, ●, ●, ●)	(↑, ↓, ↓, ↑, ●, ○, ○)
(↑, ↓, ↑, ↓, ↓, ●, ●, ●)	(↑, ↓, ↑, ↓, ○, ●, ●)
(↑, ↓, ↑, ↑, ↓, ●, ●, ●)	(↑, ↓, ↑, ↑, ○, ●, ○)
(↑, ↑, ↓, ↓, ↓, ●, ●, ●)	(↑, ↑, ↓, ↓, ●, ○, ○)
(↑, ↑, ↓, ↑, ↓, ●, ●, ●)	(↑, ↑, ↓, ↑, ○, ●, ●)
(↑, ↑, ↑, ↓, ↓, ●, ●, ●)	(↑, ↑, ↑, ↓, ○, ●, ○)
(↑, ↑, ↑, ↑, ↓, ●, ●, ●)	(↑, ↑, ↑, ↑, ○, ○, ●)

state of one of the three lamps, being either ○ ('light on') or ● ('light off'). Hence, an example of such a state would be (↑, ↓, ↓, ↑, ●, ○, ○).

We can then proceed to enumerate the state transition table, supposing that we start off in some reference state, like (↓, ↓, ↓, ↓, ●, ●, ●), and flip switches to set up an initial state S_{in} , then recording the final state S_{fin} it evolves into. Table 1 catalogues the resulting transitions.

The issue is, now, how to decide which computation—if any—the box performs. We are, in this task, not unlike somebody who, having no knowledge of the Arabic numerals, discovers a calculator—to them, keys labeled '2', '3', or even, we may suppose, '+' and '=' will carry no apparent connotation. All they could discover would be the device's reactions to certain key presses. Does this suffice to decide what computation is being performed?

The answer to this question, as argued in (Szangolies 2020), turns out to be 'no'. In particular, the three functions shown in Table 2 (alongside others) can be associated with the box on exactly equivalent grounds.

Each of these functions is simply obtained by mapping switch positions and lamp states to logical values—a mapping equivalent to the one taking the label '3' on the button of a calculator to the number 3. The function $f_A(x_1, x_2)$ is obtained by grouping two switches together—as already indicated in denoting them as s_{11}, s_{12} and so on—, then interpreting '↑' to stand for 1, '↓' as 0, '○' to denote 1, and '●' again as 0, then translating the resulting three binary values into decimal notation. Hence, using this mapping, the state (↑, ↓, ↓, ↑, ●, ○, ○) decodes to the triplet (2, 1, 3), yielding $f_A(2, 1) = 3$ (which, incidentally, means that $f_A(x_1, x_2) = x_1 + x_2$).

To obtain $f_B(x_1, x_2)$, the bit-values are simply switched, making (↑, ↓, ↓, ↑, ●, ○, ○) into (1, 2, 4), yielding $f_B(1, 2) = 4$, and for $f_C(x_1, x_2)$, the binary num-

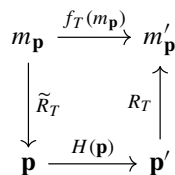
Table 2 Functions ascribed to the box

x_1	x_2	$f_A(x_1, x_2)$	$f_B(x_1, x_2)$	$f_C(x_1, x_2)$
0	0	0	1	0
0	1	1	2	2
0	2	2	3	4
0	3	3	4	6
1	0	1	2	2
1	1	2	3	1
1	2	3	4	6
1	3	4	5	5
2	0	2	3	4
2	1	3	4	6
2	2	4	5	2
2	3	5	6	1
3	0	3	4	6
3	1	4	5	5
3	2	5	6	1
3	3	6	7	3

bers are read from right to left—thus making the leftmost bit the least significant one—which yields (1, 2, 6), or $f_C(1, 2) = 6$.

To make sense of this situation, the abstraction/representation (A/R) account of computation, due to Horsman et al. (2014) (see also [Horsman 2015]) is appealed to. In brief, the account employs a special relation, called a representation relation R_T , to mediate between the physical states of a system and the abstract quantities of a computation. This relation, as indicated by the subscript T , depends on a particular theoretical model of the system in question, such that different theories T will, in general, lead to different representations, and consequently, different computations being implemented by the same physical system.

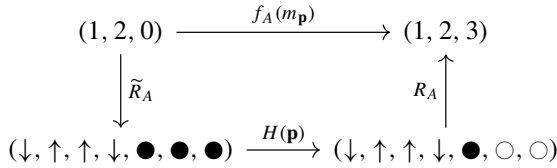
The general form of the A/R-account is given by the following diagram:



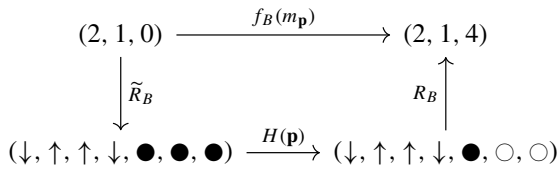
Here, \tilde{R}_T is the inverse of the representation relation, the instantiation relation, and is used to encode a certain formal object into the starting state of the system’s evolution. The system itself starts in the state \mathbf{p} , and, under physical dynamics $H(\mathbf{p})$, implemented by some Hamiltonian H , transitions into the final state \mathbf{p}' . Under the

representation relation R_T , defined by some appropriate theory T of the system, the initial physical state maps to a certain abstract representation $m_{\mathbf{p}}$, perhaps a phase-space point, or a vector in some suitable Hilbert space. The evolution $H(\mathbf{p})$ then induces a corresponding evolution $m_{\mathbf{p}} \rightarrow m_{\mathbf{p}'}$, which realizes some function taking the abstract object $m_{\mathbf{p}}$ as input and returning $f_T(m_{\mathbf{p}}) = m_{\mathbf{p}'}$ as output—this is the computation implemented by the system under the representation relation R_T .

This is best illustrated by example. The following diagram shows how the computation corresponding to f_A is carried out by the system during a certain state transition:



Under the representation relation R_A , physical states of the box are mapped to number triplets, realizing a particular computation. A change in the underlying theory—and hence, a change in representation relation—then induces a change in the computation being implemented:



The central contention of Szangolies (2020), then, is that the representation relation R_T is to be instantiated via the von Neumann process, by having the theory T being given to a Replicator \mathcal{R} as part of its tape—that is, by having this theory represented to itself.

This takes heed of the caution regarding the nature of R_T from (Horsman 2015, p. 6):

This is the core of AR theory, and it cannot be over-emphasized: *R as a representation relation is not a mathematical relation*. Neither is it a logical relation. It is a relation whose domain is physical objects and whose range is abstract/mathematical objects: it is a *representation* relation. This is an entirely new kind of relation, one that is key to understanding how physical computing devices operate by mediating between the level of physics and the level of data manipulation. [original emphasis]

It is the purpose of the following sections to further clarify the way by which the representation relation can be embodied within a von Neumann replicator, and to point out some potential obstacles and ways to overcome them.

3 Löbian Difficulties

A basic question that a replicator \mathcal{R} must be able to answer, based on its access to its own code via ‘quining’ of $\mathcal{T}_{\mathcal{N}}$, is whether to construct a certain action pattern \mathcal{A} —including whether to self-replicate ($\mathcal{A} = \mathcal{R}$) or self-modify ($\mathcal{A} = \mathcal{R}'$). What conditions need to be fulfilled to license $\mathcal{R} \rightsquigarrow \mathcal{A}$?

The question of when an agent is justified in constructing a (possibly improved) successor agent was investigated by Yudkowsky and Herreshoff (2013). There, an agent’s *criterion of action* is considered. It is assumed that there is some goal G , some state of the world such that it is in \mathcal{R} ’s interest to bring this state about; hence, \mathcal{R} is licensed only to take those actions that will aid in bringing that state about. We thus write:

$$\mathcal{R}^0 \rightsquigarrow \mathcal{A} \iff \Box_0(\mathcal{R}^0 \rightsquigarrow \mathcal{A} \longrightarrow G) \tag{3.1}$$

Here, we have introduced a generation index for the replicator; in the following, \mathcal{R}^0 will denote the current generation of a descending chain of replicators, \mathcal{R}^1 the parent, \mathcal{R}^2 the parent’s parent, and so on (the reason for the descending ordering will become obvious soon).

The box \Box_0 is the modal provability operator. If $Prv_T(x, y)$ is the provability predicate of T , which expresses that x and y are Gödel numbers such that x codes for a correct proof of the statement with Gödel number y , then $\Box_T \phi \equiv \exists x : Prv_T(x, \lceil \phi \rceil)$ (where $\lceil \phi \rceil$ denotes the Gödel number of the formula ϕ)—that is, $\Box_T \phi$ expresses that ϕ is provable in T .

Consequently, the action criterion (3.1) expresses that \mathcal{R} constructs \mathcal{A} if and only if \mathcal{R} can prove that constructing \mathcal{A} leads to the realization of G . Such proof is made possible through quining: that is, through having access to its own representation, coded in an appropriate way.

On occasion, it may be advantageous, if no \mathcal{A} is found such that G can be achieved, to not simply stay inactive, but rather, to perform actions that do not immediately lead to G , but at least, do no harm. Hence, we will allow actions to be taken that do not significantly impact the state of the world, leaving it substantially unaltered, denoted by \emptyset :

$$\mathcal{R}^0 \rightsquigarrow \mathcal{A} \iff \Box_0 \left(\mathcal{R}^0 \rightsquigarrow \mathcal{A} \longrightarrow (G \vee \emptyset) \right) \tag{3.2}$$

Such a criterion of action, however, can be problematic. For suppose that some replicator \mathcal{R}^1 wants to license the creation of a successor pattern \mathcal{R}^0 . To do so, it will have to be sure that every action taken by that successor will be directed at achieving

G—or at least, doing no harm. That is, the following must be the case:

$$\mathcal{R}^1 \rightsquigarrow \mathcal{R}^0 \iff \Box_1 \left[\forall \mathcal{A} : \mathcal{R}^0 \rightsquigarrow \mathcal{A} \iff \Box_0 \left(\mathcal{R}^0 \rightsquigarrow \mathcal{A} \longrightarrow (G \vee \emptyset) \right) \right] \tag{3.3}$$

In words, \mathcal{R}^1 constructs \mathcal{R}^0 only if \mathcal{R}^1 can prove that \mathcal{R}^0 only constructs a pattern \mathcal{A} if \mathcal{R}^0 has proven that doing so achieves G (or does nothing). We would then expect that from this, it follows that

$$\mathcal{R}^1 \rightsquigarrow \mathcal{R}^0 \longrightarrow (G \vee \emptyset), \tag{3.4}$$

thus fulfilling the criterion of action (3.2) and licensing \mathcal{R}^1 to construct \mathcal{R}^0 . However, this is unfortunately not the case.

The reason for this is Löb’s theorem (Löb 1955) (see [LaVictoire 2015] for an accessible introduction). Assume for the moment that \mathcal{R}^0 and \mathcal{R}^1 reason over the same system; then, in order to conclude that (3.4) follows from (3.3), \mathcal{R}^1 would have to believe that

$$\forall \mathcal{A} : \Box_1 \left(\mathcal{R}^1 \rightsquigarrow \mathcal{A} \longrightarrow (G \vee \emptyset) \right) \longrightarrow \mathcal{R}^1 \rightsquigarrow \mathcal{A} \longrightarrow (G \vee \emptyset). \tag{3.5}$$

This is because, to license the construction of \mathcal{R}^0 , \mathcal{R}^1 must have proven that \mathcal{R}^0 ’s reasoning is sound—that is, it must trust that whenever \mathcal{R}^0 proves that the construction of an action pattern achieves the goal (or does nothing), then this is, in fact, so. But if both \mathcal{R}^0 and \mathcal{R}^1 reason over the same system, this is equivalent to proving that all of \mathcal{R}^1 ’s own proofs that constructing an action pattern \mathcal{A} achieves $G \vee \emptyset$ obtain.

Suppose \mathcal{R}^1 could establish that all of \mathcal{R}^0 ’s proofs yield truths. Then, if \mathcal{R}^1 can prove that \mathcal{R}^0 only constructs a pattern \mathcal{A} if it has proven that doing so obtains $G \vee \emptyset$, \mathcal{R}^1 can also prove that \mathcal{R}^0 only constructs a pattern \mathcal{A} if doing so *in fact does* obtain $G \vee \emptyset$, and that consequently, constructing \mathcal{R}^0 is itself an act that works toward achieving $G \vee \emptyset$ —which is just what is asserted by Eq. (3.4).

The above is of the form $\Box \phi \longrightarrow \phi$. Löb’s theorem then asserts that whenever this can be proven, ϕ itself can be proven:

$$\Box (\Box \phi \longrightarrow \phi) \longrightarrow \Box \phi \tag{3.6}$$

Adding (3.6) to the axioms of the basic modal logic K , one obtains *provability logic* (known also as *Gödel-Löb modal logic*) GL (Boolos 1995).

Instantiating (3.6) with $\phi = \perp$ (where \perp denotes falsehood) yields:

$$\Box (\Box \perp \longrightarrow \perp) \longrightarrow \Box \perp \tag{3.7}$$

In other words, we can only prove that no falsehoods are proven by a system (i. e. that it is consistent) if it is, in fact, inconsistent (which is nothing but Gödel’s second incompleteness theorem [Gödel 1931]).

Consequently, \mathcal{R}^1 cannot in general prove (3.5), on pain of inconsistency. But then, it cannot in general license the construction of \mathcal{R}^0 . This difficulty is referred to as the *Löbian obstacle* in (Yudkowsky and Herreshoff 2013).

An ugly, but straightforward, way out of trouble is to suppose that \mathcal{R}^1 uses a strictly stronger formal system than \mathcal{R}^0 to aid its decision making. In particular, we may suppose that, if \mathcal{R}^0 reasons in some theory T^0 (which we shall assume strong enough to formalize Peano arithmetic), then \mathcal{R}^1 reasons in $T^1 = T^0 \cup \neg \Box_0 \perp$, i. e. T_0 plus the fact that T_0 is consistent, \mathcal{R}^2 reasons in $T^2 = T^1 \cup \neg \Box_1 \perp$, and so on.

Thus, each replicator must be the descendant of progressively stronger systems, increasing backwards without limit. Here, we meet the homunculus regress in its most refined form, 17th century preformationism coming back to haunt us in the austere domain of modern mathematical logic.

However, as we will see, there is cause for hope. But first, we have to discuss a striking consequence of Löb’s theorem, namely, the existence of *modal fixed points*.

3.1 Modal Fixed Points

As we have seen, using the description on the tape, a replicator \mathcal{R} can prove theorems about itself. However, what we really want is that a replicator uses itself as a representation of the state of the world—as we had surmised, in order to trigger the action ‘grab the apple of the table’, the replicator ought to have some beliefs about the world, such as that there *is* an apple on the table to grab, that grabbing that apple will achieve the goal of acquiring nutrition, and so on.

It is not immediately obvious how to turn a replicator’s self-referential nature into this sort of other-directed referentiality, or aboutness. How does the replicator interpret itself as being *about* some beliefs regarding the environment?

The answer, I want to suggest, lies in the notion of modal fixed points. Modal fixed points essentially allow us to eliminate the ‘self’ from self-referential statements, and fashion a reference to external propositions; moreover, this reference is accessible from within the system itself.

Thus, suppose we have a formula of the form

$$p \longleftrightarrow \Phi(p, q_1, \dots, q_n), \tag{3.8}$$

such that every instance of p in Φ occurs under the scope of the provability operator \Box (Φ is *modalized* in p). Then, there exists a formula $\tilde{\Phi}(q_1, \dots, q_n)$ such that

$$\Box(p \longleftrightarrow \Phi(p, q_1, \dots, q_n)) \longleftrightarrow \Box(p \longleftrightarrow \tilde{\Phi}(q_1, \dots, q_n)). \tag{3.9}$$

That is, the formula $\tilde{\Phi}$ is provably equivalent to p , and doesn’t mention p at all! This is the fixed-point theorem of *GL* (Boolos 1995, chap. 8).

For the Gödel sentence $G \longleftrightarrow \neg \Box G$, this yields

$$\Box(G \longleftrightarrow \neg \Box G) \longleftrightarrow \Box(G \longleftrightarrow \neg \Box \perp), \tag{3.10}$$

that is, the Gödel sentence is equivalent to the consistency of the system.

Now note that the criterion of action (3.2) is essentially a formula of this type:

$$\mathcal{R}^0 \rightsquigarrow \mathcal{A} \iff \Phi(R^0 \rightsquigarrow \mathcal{A}, q_i), \tag{3.11}$$

where the q_i essentially pertain to the goal G and the ‘unchanged’ world-state \emptyset . Consequently, by the fixed-point property, this is equivalent to a formula

$$\mathcal{R}^0 \rightsquigarrow \mathcal{A} \iff \tilde{\Phi}(q_i), \tag{3.12}$$

or, in other words: the license to take some action (i. e. constructing some action-pattern \mathcal{A}) is equivalent to a certain belief (or set of beliefs) about the world. In fact, for the criterion of action, the formula

$$\Box_0 \left(\mathcal{R}^0 \rightsquigarrow \mathcal{A} \implies (G \vee \emptyset) \right) \tag{3.13}$$

yields the modal fixed point¹ (Boolos 1995, p. 105, ex. 9)

$$\Box_0(G \vee \emptyset). \tag{3.14}$$

With $(G \vee \emptyset)$ referring to the state of the world after the action has been undertaken, such that the goal has been achieved (or nothing changed), this means that the intentional object of the state of mind ultimately causing some action to be performed is the state of the world brought about by the performance of said action. An action is directed at bringing about G , and so, too, is the state of mind producing that action. Hence, grabbing for the apple means that the agent believes that there is an apple on the table, which when grabbed for (and bitten into), yields the desired nutrition.

In this sense, the intentional content of a certain state of mind is given by the modal fixed point formula corresponding to the licensing of a certain action—thus, in creating an action-pattern causing the agent to grab for an apple, that agent’s intentional state is given by the fixed-point formula $\tilde{\Phi}(q_i)$, where the q_i pertain to the goal G and/or the unmodified state of the world \emptyset , as given explicitly in Eq. (3.14).

4 Subjective Experience and Intrinsic Properties

Up to this point, we seem not to have made much headway. While it is intriguing that the preceding discussion seems to have opened up the possibility of fashioning an account of intentionality by means of modal fixed points, it comes at the expense of introducing an infinite regress of formal systems, with each parent generation needing to be more powerful than its successor.

This difficulty was, in slightly different terms, already noted in (Szangolies 2020). There, it was argued that the representation relation R_T , the central object of the A/R-theory, cannot itself be computational. The reason for this is just the above regress: if representation⁰ itself is a computation, then we would need a representation¹ relation to

¹ I thank an anonymous referee for drawing my attention to this point.

implement it; and again, a representation² relation to implement this one; and so on. In particular, this means that an agent cannot simply instantiate a computation by running some quining program, as whether that program is computed is itself only settled by an appropriate implementation relation. Hence, the caution of Horsman (2015) that the representation relation cannot be a mathematical, or indeed, logical/computational relation, is seen to be perfectly apt.

But how, then, to ground the regress and furnish a definite implementation of the computation performed by our box? A hint is provided by the fact that, as observed by Copeland (1996), the trivialisation challenge to computational implementation is essentially the objection raised by Newman (1928) against Bertrand Russell's structuralism (Russell 1927).

Russell had argued that all we can ever infer from perception is structure (in the sense of relational structure). But then, as Newman pointed out, any collection of elements can be considered to support any arbitrary structure, as long as there are sufficiently many of them. All we could empirically discover, on such a conception, would then be answers to questions of cardinality.

This, then, suggests that the answer to the problem of implementation, to how to ground the regress, is the same as the answer to Newman's objection.

4.1 Monism and Models: Seager's Analogy

An answer to the problem posed by Newman is often considered to be that, in some sense, while we can only grasp the external world in terms of structure, we are, in experience, directly aware of non-structural elements of the world; of intrinsic properties, which are brought to bear the structure of the world.

Seager (2016) glosses this in terms of an intriguing analogy. Considering the Gödel sentence G , he notes (Seager 2016, p. 337):

[T]here must be a model of the axioms of arithmetic that make G true. And so there is. But, equally, there must also be a model of the axioms that makes G come out false. Else, by the completeness of first order logic, G would be provable from the axioms. [...]

The analogy with [...] the Newman problem is that [...] science cannot specify more than the bare structure of the world, rather in the way the axioms of arithmetic cannot, obviously, specify more than the structure which is compatible with all its models. The analogue of the Newman objection would be for someone to complain that the axioms of arithmetic do not by themselves fix the domain of interpretation. [...] The answer to the arithmetical analogue-Newman problem is that we have a grasp of the mathematical domain at issue that goes beyond the structure imposed by the axioms. Similarly, to answer the Newman objection in the case of science, there must be some way for us to grasp the domain to which scientific theory applies which goes beyond the merely structural or relational constraints imposed by the formal structure (mathematical structure) of scientific theories.

What is thus needed, in other words, to solve the Newman problem, is the specification of a model of the ‘axioms’ that define our theories of the external world. By Copeland’s argumentation, we may propose that this will also yield a sufficient answer to the problem of implementation.

This proposal was raised also in (Szangolies 2020). With the apparatus developed here so far, we have the tools needed to give it a more explicit formal realization.

4.2 Self-modification: The Impact of the World Without

As we have seen previously, in the general case, licensing the construction of successor patterns runs into difficulties of self-reference. However, it also suggests a way to turn self-reference into ‘other-reference’, to turn navel-gazing introspection into beliefs directed towards the rest of the world, by means of modal fixed points.

Taking Seager’s analogy seriously, these difficulties stem from Newman’s problem—mere relation does not suffice to settle anything safe questions of cardinality. We must thus go beyond structure—specify a concrete model realizing that structure.

When in contact with the exterior world, a replicator may need to construct successors having access to data it cannot accurately predict—which says nothing but that external influences may lead to a change of mind, so to speak. New data will yield new beliefs about the world, and successive mind-states, replicators, need to be able to reason soundly given new data; hence, any given replicator must be sure that any successor it creates will, given true data, come only to true conclusions about which actions will precipitate the achievement of a certain goal—without, however, having access to this data.

This is a more complex situation than we faced previously: \mathcal{R}^1 needs not merely to trust in the reasoning of \mathcal{R}^0 , but needs to trust \mathcal{R}^0 even if it has been modified by the addition of arbitrary (true) statements about the environment. Following Yudkowsky and Herreshoff (2013), we can model this by adjoining an explicit model τ of ‘the territory’, encoding the environment. To a replicator \mathcal{R}^0 , this will be presented in terms of its tape $\mathcal{T}_{\mathcal{N}}$. This will necessitate that \mathcal{R}^1 be able to reason about things ‘true in the territory’, which requires it to be able to reason within a system strong enough to be able to formalize statements of the form ‘ $\tau \models [\phi]$ ’, i. e. ‘the (quoted) formula ϕ is true within the (quoted) model τ ’. Zermelo-Fraenkel set theory (ZF) has the requisite expressive power (Yudkowsky and Herreshoff 2013).

The model τ , in general, consists of sets denoting a certain domain and the relations over this domain—for example, for Peano arithmetic, it would contain a set of elements—namely, the natural numbers—, a set of ordered pairs modeling the successor relation (e. g. (3, 4)), a set of ordered triples for addition—(2, 3, 5), and so on—and one for multiplication—(2, 4, 8), etc., thus yielding the standard model of Peano arithmetic.

In our case, one salient part of the ‘territory’ is given by the box and the computation it implements. We can think of this as being given by some pattern \mathcal{E} on the tape $\mathcal{T}_{\mathcal{N}}$, induced by environmental influences (say, the inspection of the box by the agent). This has the following ingredients:

1. A set $\{\uparrow, \downarrow, \circ, \bullet\}$ to describe the physical state of the system

2. A set $\{0, 1\}$ to describe the abstract representation of the system
3. A relation consisting of the ordered 7-tuples $\{(\downarrow, \downarrow, \downarrow, \downarrow, \bullet, \bullet, \bullet), (\downarrow, \downarrow, \downarrow, \uparrow, \bullet, \bullet, \circ), \dots, (\uparrow, \uparrow, \uparrow, \uparrow, \circ, \circ, \bullet)\}$ describing the physical behavior of the system
4. A relation consisting of the triples $\{(0, 0, 0), (0, 1, 1), (1, 0, 1), (1, 1, 0)\}$
5. A relation consisting of the triples $\{(0, 0, 0), (0, 1, 0), (1, 0, 0), (1, 1, 1)\}$
6. A relation consisting of the tuples $\{(\uparrow, 1), (\downarrow, 0), (\circ, 1), (\bullet, 0)\}$
7. A relation consisting of the tuples $\{(1, \uparrow), (0, \downarrow), (1, \circ), (0, \bullet)\}$

1 and 2 simply enumerate the objects of our universe—both physical and abstract. 3 yields all allowed physical configurations of the system, thus characterizing the physical state space. 4 and 5 then essentially characterize the elements 1 and 0: they represent, respectively, addition and multiplication over the finite field of two elements, $GF(2)$. With these relations, one readily checks that $\{1, 0\}$ fulfills the field axioms F —multiplication and addition are both associative and commutative, there exists an identity element (1 resp. 0) for either, every element has an inverse for both operations (except for 0 in the multiplicative case), and multiplication distributes over addition.

Finally, the last two relations, 6 and 7, tell us the connection between the physical states and the abstract symbols. Together 1 through 7 form a model of the theory T used to furnish the representation relation R_T —a (presumed to be) complete (to some degree of accuracy) physical description of a system, together with the description of an abstract space within which its states are to be represented, and the mapping between these.

At first glance, it might seem odd to consider the abstract representations of physical states an element of a physical theory. In general, theories, as they are usually specified, include only some reference to the abstract arena within which systems are represented (the theory’s state space)—in quantum mechanics, for instance, each system is associated with a separable complex Hilbert space equipped with an appropriate inner product, and states of the system are represented by rays (subspaces of complex dimension 1) within that Hilbert space.

This is less specific than the above: no prescription is given as to which state to associate with what ray. However, in applying the theory to any given system, such a choice has to be made: for instance, an electron’s ‘spin up’-state may be associated with the state $|1\rangle = (0, 1)^T$. The label ‘spin up’ itself then may be thought to be experimentally individuated—say, as that state the electron is in if it is deflected upwards in a Stern-Gerlach type experiment. It is then this theory as specified to a concrete system that we mean when speaking of ‘theory’ generally.

The example of the field axioms shows that in choosing this particular model, we might have had alternatives—for instance, by the use of a different field, a different realization of the field axioms, but more importantly, just as well by using a different choice for the mapping between physical and abstract objects. Thus, substituting 6 and 7 by

- 6'. A relation consisting of the tuples $\{(\uparrow, 0), (\downarrow, 1), (\circ, 0), (\bullet, 1)\}$
- 7'. A relation consisting of the tuples $\{(0, \uparrow), (1, \downarrow), (0, \circ), (1, \bullet)\}$,

an equivalent concrete realization of the same abstract structure would be achieved. However, using the representation relation generated by this particular model, the

box is seen to compute $f_B(x_1, x_2)$ instead of $f_A(x_1, x_2)$. Consequently, the difference between the implemented computations corresponds to the difference between the models of the same formal description of the system—different models of the same axioms, in this sense.

Using its tape, a replicator \mathcal{R}^1 can then reason about what \mathcal{R}^0 would conclude, given some state of the environment. To do so, \mathcal{R}^1 must be capable of proving that every formula proven by a successor replicator \mathcal{R}^0 about the environment must, in fact, be true of it; but this is just what is achieved by augmenting \mathcal{R}^1 with an explicit model τ of the environment, and assuming that its reasoning capabilities are strong enough to represent the notion of semantic entailment (hence, the need to appeal to ZF). Essentially, augmenting \mathcal{R}^1 by τ has the same effect as having it reason within a formal system strictly stronger than that of \mathcal{R}^0 . Including it serves as a reminder that the mind, as given by the replicator \mathcal{R}^1 , is not a merely formal object, but possessed of structure-transcending as well as structural properties.

This allows us to overcome the *Löbian obstacle*, enabling \mathcal{R}^1 to assert

$$\forall \phi : \Box_1 [(\tau \models \Box_0 \phi) \longrightarrow (\tau \models \phi)]. \tag{4.1}$$

This is again of the general form ‘ $\Box \Phi \longrightarrow \Phi$ ’, that is, it licenses \mathcal{R}^1 to conclude that whatever is proven by \mathcal{R}^0 , in fact obtains (in the quoted model τ).

With $\phi = \mathcal{R}^0 \rightsquigarrow \mathcal{A} \longrightarrow (G \vee \emptyset)$, this provides the replacement of (3.5). Consequently, if \mathcal{R}^0 can establish this, it can establish that \mathcal{R}^0 only constructs an action pattern if doing so achieves $G \vee \emptyset$, and that hence, constructing \mathcal{R}^0 is itself an action achieving $G \vee \emptyset$. In result, the criterion of action (3.2) can be fulfilled, and construction of \mathcal{R}^0 will be licensed.

Thus, \mathcal{R}^1 only constructs \mathcal{R}^0 , if it can prove that within the quoted model τ , whenever \mathcal{R}^0 proves something, then it is in fact true within τ . Note the implicit double quoting: $\Box_0 \phi$ denotes the \mathcal{R}^1 -Gödel number of $\exists x : Prv_0(x, \lfloor \phi \rfloor)$, where $\lfloor \phi \rfloor$ denotes the \mathcal{R}^0 -Gödel number of ϕ —that is, \mathcal{R}^1 keeps an internal representation of \mathcal{R}^0 's representation of arbitrary formulae (Yudkowsky and Herreshoff 2013). In other words, \mathcal{R}^1 will have beliefs about the beliefs of \mathcal{R}^0 —in the case where both are equivalent, then, beliefs about its own beliefs. In some sense, a replicator must represent its own beliefs to itself.

In the end, this is nothing but an explicit version of (3.5), viewed from the perspective of a stronger formal system than that used by \mathcal{R}^0 . If \mathcal{R}^0 reasons within T^0 , we can take this system to be $T^1 = T^0 \cup \neg \Box_0 \perp \text{---} T^0$ ‘enhanced’ by the proposition that T^0 is consistent. This is equivalent to specifying a model τ of T^0 ; within that model, \mathcal{R}^1 can then conclude that all of \mathcal{R}^0 's proofs yield truths, thus overcoming the Löbian difficulty. Hence, importantly, having access to a model of a given system is equivalent to moving to a stronger system; using Seager’s analogy, intrinsic properties then outstrip what can be formally captured by means of axiomatic reasoning.

This illustrates the importance of the self-referential nature of von Neumann’s construction. Given a model of itself (or a successor pattern) together with the environment, a replicator \mathcal{R} is capable of reasoning about its own actions within the environment. Moreover, differences in the model—as in, different models of the same axiomatic structure being supplied to the replicator—will result in different actions: the criterion

of action (3.2) will license different actions, given differences in the model. Consider, for instance, the emission of an action pattern causing the agent to flip a switch of the box: depending on whether the switch-up state \uparrow is mapped to 0 or 1, reaching a certain goal—say, to represent the value 1—will be facilitated by either flipping, or not flipping.

For a replicator $\mathcal{R} = \mathcal{N} + \mathcal{T}_{\mathcal{N}}$, the notation ‘ $\tau \models$ ’ in (4.1) must then be understood as a formalization having access to its own intrinsic properties—ultimately, via quining, or rather, using the tape $\mathcal{T}_{\mathcal{N}}$ to talk about the entire assembly \mathcal{R} , including an encoded representation of the external world. This will not, of course, enable it to prove itself consistent. In appealing to an infinite regress of progressively stronger formal systems, we are, essentially, confusing the map for the territory: it is not important what \mathcal{R} can prove about itself, but rather, what is *true* about \mathcal{R} .

Consider again the modal fixed point corresponding to the criterion of action (3.2). We know that \mathcal{R}^1 can prove the following:

$$\mathcal{R}^1 \rightsquigarrow \mathcal{A} \iff \tilde{\Phi}(q_i) \tag{4.2}$$

That is, licensing the action of constructing \mathcal{A} is equivalent to a certain belief $\tilde{\Phi}(q_i)$ about the world—setting up the excitation pattern in the motor cortex that leads to grabbing an apple and biting into it is equivalent to believing that there is an apple at such-and-such a position, and that apples yield nutrition. The truth of the fixed-point formula itself is then independent of the formal system expressing \mathcal{R}^1 ’s reasoning capabilities; indeed, the above discussion implies that, for any \mathcal{A} equivalent to some \mathcal{R}^0 of equal capacities to \mathcal{R}^1 , it must be equivalent to an assertion of that system’s consistency.

However, in every given model, it will either be true, or false. In particular, it will either be true, or false, of \mathcal{R}^1 , itself— \mathcal{R}^1 either has, or does not have this particular belief. But then, the having of that particular belief, while a fact that cannot be derived from any formal specification of \mathcal{R}^1 , leapfrogs the need to appeal to a stronger system. We do not have to establish the formula $\Phi(\mathcal{R}^1 \rightsquigarrow \mathcal{R}^0, q_i)$, i. e.

$$\Box_1 \left(\mathcal{R}^1 \rightsquigarrow \mathcal{R}^0 \implies (G \vee \emptyset) \right), \tag{4.3}$$

and thus, need not appeal to (4.1), but only need $\tilde{\Phi}(q_i)$ instead.

This should not be misunderstood to imply that \mathcal{R} can, in some mysterious way, establish the *truth* of $\tilde{\Phi}(q_i)$, despite it not being provable within the system of reasoning used by \mathcal{R} . The point is, rather, that \mathcal{R} does not, in general, *need* to establish its truth. The proposition $\tilde{\Phi}(q_i)$ can be considered to be of the form ‘ \mathcal{R} believes that q_i ’, which is true if \mathcal{R} believes that q_i , without there being any need for \mathcal{R} to believe that (read: have proven that) ‘ \mathcal{R} believes that q_i ’. Rather, $\tilde{\Phi}(q_i)$ stands to \mathcal{R} ’s beliefs as the proposition ‘the car is red’ to the car actually being red: there is no need for the car to prove that it is red in order for it to have that color.

Hence, this is not an argument that the mind has powers exceeding that of any formal system, as argued most famously by Lucas (1961) and Penrose (1989), but also Gödel

(1995) himself; on the contrary, its reasoning capacities are explicitly bounded by what \mathcal{R} can prove about itself.

One might claim that an agent always knows that it believes what it believes. But this is not generally the case: many of our actions imply beliefs that may never have consciously occurred to us. Indeed, it is in general a nontrivial matter to work out what beliefs a certain stance commits us to. There is no need to know that we believe that apples are nutritious to eat them.

However, knowledge of (at least some of) our beliefs does come about through this mechanism. The reason for this is the double-quoting of 4.1: in general, a replicator will have beliefs about its successor's beliefs in order to license its construction.

While it is thus impossible to establish a license for the construction of a successor pattern by formal means, introspecting on \mathcal{R}^1 's beliefs suffices to avert the detour into having to prove a potentially infinite chain of successor patterns sound.

4.3 Introspecting Intrinsic Properties

The preceding discussion seems to make introspection a deeply mysterious capability. Indeed, even to \mathcal{R} itself, its introspective knowledge will remain fundamentally unanalysable: if pressed, it could not give any sort of reason for its judgment that its successor's reasoning will be sound—as any such reasoning would yield a consistency proof, and with the successor's reasoning capacities equalling its own, hence imply inconsistency.

But we should not be surprised by this essential boundary to analysis: after all, we are trying to formulate a theory of something that, if Russell's structural approach to science is apt, is fundamentally beyond any theory's grasp—the intrinsic, structure-transcending (Strawson 2019) properties that ground the structure of the world as relayed to us in our theories. We are thus faced with something that cannot be formalized within any set of axioms, nor be the result of any computation whatsoever—and that precisely because of this, has the right properties to fill in the concrete character missing from mere structural specification, and, as discussed above, provide an answer to Newman's objection, the homunculus problem, and the question of computational implementation—which are thus revealed to be nothing but facets of the same issue.

The intrinsic properties, as elaborated, then must be something inaccessible to formal reasoning—something deeply mysterious, yet present to introspective examination. It is then a natural leap to identify them with the elements of subjective experience, with qualia—as indeed proposed in (Szangolies 2020). After all, whatever structure the world has is, in experience, individuated by subjective qualities of the objects of our experience.

Take, for instance, the structure of the visual field, as borne by the visual qualities of the objects within it: consider a uniformly blue wall, illuminated by a single light source, say a bulb some distance d removed from the wall's center. There will then be a relation 'brighter than' that holds between any two points a and b on the wall when a is closer to the illuminated center than b . This will partition the wall into concentric circles of equal brightness. It is then the *subjective impression* of brightness that

presents this structure to us, and singles it out among all the other relations fulfilled by points of the wall.

The proposal presented herein is then, essentially, that this singling-out of the relational structure fulfilled by points on the wall proceeds by means of intrinsic properties being 'brought to attention' via the von Neumann process—in experience, we are directly presented with a concrete instantiation of the 'brightness' relation, just as an \mathcal{R}^1 -replicator is presented with a concrete instantiation of the structural relations between itself and its environment, as embodied in its tape.

While we thus cannot say much about the precise mechanism by which the intrinsic properties become present in experience, its *means* seems quite clear: as discussed above, the criterion of action (3.2) is a self-referential formula giving rise to a modal fixed point, and hence, to beliefs about the world. Moreover, though, its truth value—and thereby, the having of the attendant beliefs, which in general include meta-beliefs about beliefs—is, like that of the Gödel sentence, dependent on the model of the formal system—and thus, with Seager's analogy, on the intrinsic properties. Hence, via self-reference, intrinsic properties enter as determining conditions into our beliefs about the world. Modal fixed points are structural objects dependent on intrinsic properties, thus bringing the latter within the purview of the former.

In particular, the action of creating a new replicator—a new state of mind—is either licensed or prohibited by the intrinsic properties; in this way, a replicator looking to create a successor will do so depending on the underlying model, thus eliminating the need of a consistency proof in judging whether to carry out the construction. In this sense, the action of creating a successor pattern must then be, as a modal fixed point, equal to an assertion of the consistency of the system—similar to the Gödel sentence. Indeed, both von Neumann's replicators and Gödel's construction share an obvious structural similarity.

This also takes care of a looming worry: what we are conscious of does not seem to pertain, in any form, to the construction of mental patterns, or even states of mind; we are, typically, conscious of the world, or take ourselves to be. But what we are conscious of is determined by the fixed point formula, which is explicitly independent of the type of action taken by the replicator.

This idea has several intriguing consequences. First of all, it straightforwardly gives a reason for the hardness of the hard problem—indeed, it shows it to be not merely hard, but impossible to solve: intrinsic properties cannot be reduced to structural notions, just as the truth of the Gödel sentence does not follow from a given set of axioms. We are cognitively closed (Kriegel 2003) with respect to these intrinsic properties—which does not signal a failure of our own particular reasoning capacities, but a consequence of the preconditions of reasoning itself: reasoning is, ultimately, instantiating the structure of one domain within another—making a computer model of something—or indeed, a mental, or physical model. That which 'does the work' of instantiating structure—the structure-transcending—then necessarily lies beyond this capacity.

Furthermore, it gives subjective experience, qualia, a respectable job to do: inferences of the form of (3.2) will come out true or false, based on the intrinsic properties—based on the model—and hence, license different actions. Qualia are

not epiphenomena; however, their efficaciousness necessarily lies beyond theoretical modeling.

5 Conclusion

The model introduced in (Szangolies 2015, 2018, 2020) and refined herein consists of a range of interlocking components from different areas of study, whose precise relationships can be difficult to disentangle. Thus, as an aid to the better appreciation of the overall structure of the argument so far, I will highlight and discuss each of these in turn.

- (1) *The von Neumann Mind*. The self-representational and self-referential qualities of von Neumann's construction enable it to collapse the homunculus regress, by eliminating the distinction between a (mental) representation and its user.

There is a widespread intuition that self-reference, or self-representation, in some form or another, might be the linchpin in explaining the emergence of mind from matter (Kriegel and Williford 2006). One approach appeals to the particular self-representational properties of von Neumann's formalism for self-reproducing automata to imbue a material base with the sort of representational characteristics present in conscious experience (Szangolies 2015; Pattee and Rączaszek-Leonardi 2012).

The chief focus of the present considerations is to throw the means by which self-reference and self-representation provide the prerequisites to mental phenomena into sharper relief. The key element of this is what Pattee (2008) terms 'matter-symbol complementarity', or in other words, the capacity of the tape of a von Neumann replicator to serve both syntactic and semantic roles—to be copied as a string of inert signs, and to be interpreted as a blueprint of some pattern to be constructed.

- (2) *Quining*. Via the tape, a von Neumann replicator obtains access to its own properties—they are represented to it, and thus, can guide action.

With the tape fulfilling this dual role, replicators, using the quining-trick discussed in Sect. 1, gain access to their own properties—essentially, they embody self-referential propositions of the form

$$p \longleftrightarrow \Phi(p, q_i),$$

where p in $\Phi(p, q_i)$ only occurs modalized, that is, under the scope of the provability operator \Box . The most familiar such proposition is the Gödel sentence

$$G \longleftrightarrow \neg\Box G,$$

asserting its own unprovability.

In particular, for any replicator \mathcal{R} , whether it licenses the construction of a successor pattern is given by a proposition of the above form: to take any action, it must first prove that doing so will bring about some goal G ; thus, in particular, to construct a

successor pattern, \mathcal{R} must prove that whatever action this successor takes will likewise further achievement of G (or at least, do no harm).

This, however, spells trouble, due to Löb’s theorem: in general, since \mathcal{R} can only trust those proofs it has actually found, and not issue a blanket approval of the form ‘whenever I prove something, it is true’, it cannot conclude that, for its successor, once it has proven some action to achieve G , this action will, in fact, achieve G . Hence, it seems that \mathcal{R} can never generally approve the construction of a successor, as achievement of G cannot be certified.

- (3) *Modal fixed points.* Self-reference can be turned into outward-directed reference by exploiting the fact that self-referential formulas can be rewritten with the self-reference eliminated.

However, a possible solution is posed by the fixed-point theorem of Gödel-Löb modal logic GL , which demonstrates the existence of modal fixed points, that is, propositions $\tilde{\Phi}(q_i)$ such that

$$\Box(p \leftrightarrow \Phi(p, q_i)) \leftrightarrow \Box(p \leftrightarrow \tilde{\Phi}(q_i)) .$$

In other words, propositions equivalent to p such that they do not mention p at all. For the Gödel sentence G , this is simply the proposition $\neg\Box\perp$, i. e. the proposition asserting the consistency of the system. In this sense, G can be considered to represent the consistency of the system via ‘quining’, i. e. referring to its own representation.

If the proposition p now pertains to the performance of some action (such as the construction of a successor), we may consider $\tilde{\Phi}(q_i)$ to be a representation of the replicator’s beliefs about the world that license this action (or fail to, as the case may be). In the concrete case of the criterion of action, Eq. (3.1), the modal fixed point is nothing but $\Box_0(G \vee \emptyset)$, the state of the world the action is directed at bringing about. That is, the action of grabbing for the apple on the table is licensed by the belief that there is such an apple, and that apples provide nutrition, and that hence, doing so will achieve the goal of acquiring nutrition.

- (4) *The intrinsic/structural dichotomy.* Propositions undecidable by the axioms of a formal system are either true or false in a given model; with the analogy between axioms/model and structure/intrinsic properties, a system’s intrinsic properties can license action even if the system cannot formally prove this.

For any model of a system, such a proposition will either be true, or false; and in general, the system will not be able to decide which. Using an analogy proposed by Seager (2016), we may think of this as a proposition made true (or false) by the intrinsic (or structure-transcending Strawson 2019) properties of a system not captured by its structural description. But as we surmised, it is equivalent to the action-licensing proposition p ; hence, whether an action is licensed depends on the replicator’s intrinsic properties. To license replication, \mathcal{R} thus never has to perform the impossible task of proving that all its (or its successors’) proofs yield truths; the truth of $\tilde{\Phi}(q_i)$, which is independent of p , suffices.

- (5) *The Hard Problem’s hardness.* Intrinsic properties are beyond formal describability; hence, just as the axioms of PA are consistent with either the truth or falsity of

the Gödel sentence, they cannot be captured by any structural account (i. e., any theory).

In the above sense, the intrinsic properties can break the deadlock of regress—at the cost of evading formal describability. This, it was proposed in (Szangolies 2020), is the origin of the ‘Hard Problem’ (Chalmers 1995) of consciousness: any formal description is consistent with either value for the structure-transcending properties, as any physical description is consistent with different sorts of phenomenal experience (as in inverted spectra [Shoemaker 1982]), or even, their utter absence (as in zombies [Chalmers 1996]).

(6) *Overcoming Newman’s problem.* Intrinsic properties provide the relata that fulfill the relations captured by formal descriptions.

Intrinsic properties, as proposed in (Szangolies 2020), are ideally suited to provide the theory T of a physical system that furnishes its computational interpretation by means of the representation relation R_T of Horsman (2015). More generally, they provide the ‘clay’ used by an agent to furnish its models of the world—that is, they allow to answer the objection of Newman (1928), levied against the structuralism of Russell (1927). Such models are present to us in subjective experience. The key point of the present proposal as regards subjective experience then is that qualia are intrinsic properties shackled to perform representational duties by the von Neumann process.

The combination of (1)–(6) then constitutes an approach to the philosophy of mind that combines elements from familiar proposals, extending them with some novel ideas regarding the origin of intentionality and the role of phenomenal/intrinsic properties.

As such, this approach is a kind of nonreductive physicalism that shares certain ground with mysterianism (McGinn 1989), Russellian or intrinsic-property monism (Alter and Nagasawa 2015), and higher-order theories (HOT) (Rosenthal 2005). Like mysterianism, according to (5), it proposes that the Hard Problem remains unsolvable: intrinsic properties are ineffable, because the effable is limited to the structural. With Russellian monism, (4) entails that it shares a commitment to properties that go beyond the structural—without, however, entailing a commitment to an inherently experiential or mental character for the intrinsic properties, hence, not leading to some kind of panpsychism or panexperientialism (Strawson 2019; Seager 2006). For this, it appeals to a mechanism to ‘bring to light’ the intrinsic properties—the von Neumann process, (1), considered as a means to imbue symbols with meaning (Szangolies 2015; Pattee 1969). This is then similar to the commitment of higher-order theories, in which a thought is conscious if it is the object of a higher-order thought—however, the self-referential nature of the von Neumann construction allows to identify base- and higher-order thought. In this, it shares characteristics with self-representational theories of consciousness (Kriegel 2003) and the ‘quotational’ higher-order theory of Picciuto (2013).

The most significant novel aspect of this proposal is, then, that von Neumann Minds, in a way similar to the notion of quining (2), can gain access to their own properties. This allows for two related advancements: first, by appealing to modal fixed points (3), the intentional content of the state of mind leading to the performance of a certain action can be brought to the fore. These modal fixed points are independent of a given replicator’s formal reasoning capabilities; hence, it is the intrinsic properties

that fix their truth-value, and thus, ultimately license actions. This is then what fixes the reference of representations, and thus, overcomes structural underdetermination, leading to an answer to Newman's objection (6).

Should the above considerations bear fruit, systems implementing processes formally equivalent to von Neumann's replicator should exist in the brain. One line of further research is then to investigate in what form neural networks—either natural or artificial—may support processes of replication. Indeed, von Neumann (1966) took the neuron model of McCulloch and Pitts (1943) as a means to implement the logical operations performed by his automaton. However, the implementation of something like the present proposal would take place entirely within a neural network. Along such lines, various proposals for self-referential, 'quining' neural networks have been formulated (Schmidhuber 1993a, b; Chang and Lipson 2018); whether they can be appealed to as a realization of the von Neumann process remains to be discussed.

In a more directly biological setting, one might investigate the role of *reentry* (Edelman and Gally 2013): the bi-directional signaling along reciprocal connections between distinct brain areas. Mumford (1991, 1992) has introduced the notion of 'active blackboards': cortical areas carry out computations with the aid of thalamic nuclei with which they are reciprocally connected. The hope is then that in this bi-partite structure, one might find reflected the relationship between a von Neumann automaton and its tape.

A final question that we can only raise here is that of how distinct concepts may come to be bound together within one unified conscious field of experience. A single replicator, one might propose, corresponds to some appropriate 'simple' concept, with an agent's state of mind being made up of a simultaneous population of such entities. What binds them together into a unified whole?

One should take care to distinguish this issue from the combination problem of panpsychism (Seager 1995): on the present proposal, there is no need to unify distinct elements of experience, associated, for instance, with individual intrinsic properties, into one larger consciousness, as conscious experience tout court only emerges upon the unification of intrinsic properties within a von Neumann replicator. However, it seems implausible (although logically possible) to have the state of mind of an agent given by a single replicator; economy, if nothing else, seems to suggest separate replicators for separate concepts, or thoughts, or whatever else the basic elements of experience might be considered to be.

One proposal might be to look towards structures that achieve self-reference only by referring to one another, such as the following pair of sentences:

- (A) Sentence B is true
- (B) Sentence A is false

Indeed, it is possible to construct simultaneous fixed points for agents accessing each other's source code, for instance, to decide whether to collaborate or defect in a multi-agent prisoner's dilemma (LaVictoire et al. 2014; Barasz et al. 2014). However, we will leave the investigation of such possibilities for future work.

Appendix A: A Self-Inspecting Quine

```

function Quine2 {
$start = "@"'
$end = "'@"

$quine = ""

$sep = [char]59
$rep = [char]63

$stape = @'
function Quine2 {
$start = "@"'
$end = "'@"

$quine = ""

$sep = [char]59
$rep = [char]63

$stape = ;

# If there is a question mark in the code, reproduction will take
# place
# ?

# Construct the first lines up to $stape = ...
$quine += $stape.Split($sep)[0] + $start + "n"

# Copy the tape
$quine += $stape

# Construct the final lines
$quine += "n" + $end + $stape.Split($sep)[1]

Write-Host "The length of this program is: $($quine.Length)"

if ($quine.Contains($rep)){
Write-Host "This program will produce a copy of itself."
Return $quine
} else {
Write-Host "This program will not produce a copy of itself."
Return $null
}
}
'@

# If there is a question mark in the code, reproduction will take
# place
# ?

# Construct the first lines up to $stape = ...
$quine += $stape.Split($sep)[0] + $start + "n"

# Copy the tape
$quine += $stape

# Construct the final lines
$quine += "n" + $end + $stape.Split($sep)[1]

Write-Host "The length of this program is: $($quine.Length)"

if ($quine.Contains($rep)){
Write-Host "This program will produce a copy of itself."
Return $quine
} else {
Write-Host "This program will not produce a copy of itself."
Return $null
}
}
}

```

Listing 2 A quine which inspects its own code, and modifies its actions based on that.

References

- Alter, T., & Nagasawa, Y. (2015). *Consciousness in the physical world: Perspectives on Russellian monism*. Oxford University Press.
- Barasz, M., Christiano, P., Fallenstein, B., Herreshoff, M., LaVictoire, P., & Yudkowsky, E. (2014). Robust cooperation in the Prisoner's Dilemma: Program equilibrium via provability logic. [arXiv:1401.5577](https://arxiv.org/abs/1401.5577) <http://arxiv.org/abs/1401.5577>
- Boolos, G. (1995). *The logic of provability*. Cambridge: Cambridge University Press.
- Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press.
- Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3), 200–219.
- Chang, O., & Lipson, H. (2018) Neural network quine. In: *Artificial life conference proceedings* (pp. 234–241). MIT Press.
- Copeland, B. J. (1996). What is computation? *Synthese*, 108(3), 335–359.
- Dennett, D. C. (1988). Quining qualia. In: *Consciousness in modern science*. Oxford University Press.
- Edelman, G. M., & Gally, J. A. (2013). Reentry: A key mechanism for integration of brain function. *Frontiers in Integrative Neuroscience*, 7, 63.
- Gödel, K. (1995). Some basic theorems on the foundations of mathematics and their implications (Gibbs Lecture, 1951). In: W. Goldfarb, C. Parsons, S. Feferman, J. W. Dawson Jr., & R. N. Solovay (Eds.), *Collected works* (Vol. III, pp. 304–323). New York: Unpublished Essays and Lectures, Oxford University Press.
- Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, 38(1), 173–198.
- Godfrey-Smith, P. (2009). Triviality arguments against functionalism. *Philosophical Studies*, 145(2), 273–295.
- Hofstadter, D. R. (2007). *I am a strange loop*. Basic books.
- Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An eternal golden braid*. Hassocks, Sussex: Harvester Press.
- Horsman, D. C. (2015). Abstraction/representation theory for heterotic physical computing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 373(2046), 20140224.
- Horsman, C., Stepney, S., Wagner, R. C., & Kendon, V. (2014). When does a physical system compute? *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 470(2169), 20140182.
- Kriegel, U., & Williford, K. (eds) (2006) *Self-representational approaches to consciousness*. MIT Press.
- Kriegel, U. (2003). The new mysterianism and the thesis of cognitive closure. *Acta Analytica*, 18(30–31), 177–191.
- Laing, R. (1977). Automaton models of reproduction by self-inspection. *Journal of Theoretical Biology*, 66(3), 437–456.
- LaVictoire, P. (2015). An introduction to Löb's theorem in MIRI research. In: *Technical reports*. Machine Intelligence Research Institute, Berkeley, CA. <http://intelligence.org/files/lob-notes-IAFF.pdf>
- LaVictoire, P., Fallenstein, B., Yudkowsky, E., Barasz, M., Christiano, P., & Herreshoff, M. (2014). Program equilibrium in the Prisoner's Dilemma via Löb's theorem. In: *Workshops at the twenty-eighth AAAI conference on artificial intelligence*.
- Löb, M. H. (1955). Solution of a problem of Leon Henkin I. *The Journal of Symbolic Logic*, 20(2), 115–118.
- Locke, J. (1690). An essay concerning humane understanding. T. Basset, E. Mory, London.
- Lucas, J. R. (1961). Minds, machines and Gödel. *Philosophy*, 36(137), 112–127.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), 115–133.
- McGinn, C. (1989). Can we solve the mind-body problem? *Mind*, 98(391), 349–366.
- Mumford, D. (1991). On the computational architecture of the Neocortex I: the role of the Thalamo-Cortical loop. *Biological Cybernetics*, 65(2), 135–145.
- Mumford, D. (1992). On the computational architecture of the Neocortex II: The role of Cortico-Cortical loops. *Biological Cybernetics*, 66(3), 241–251.
- Newman, M. H. (1928). Mr. Russell's "causal theory of perception". *Mind* 37(146), 137–148.
- Pattee, H. H., & Rączaszek-Leonardi, J. (2012). *Laws, language and life: Howard Pattee's classic papers on the physics of symbols with contemporary commentary* (vol. 7). Springer Science & Business Media.

- Pattee, H. H. (1969). How does a Molecule become a message? *Communication in Development*, 3, 1–16.
- Pattee, H. H. (2008). The necessity of biosemiotics: Matter-symbol complementarity. In M. Barbieri (Ed.), *Introduction to biosemiotics* (pp. 115–132). Berlin: Springer.
- Penrose, R. (1989). *The emperor's new mind: Concerning computers, minds, and the laws of physics*. Oxford University Press.
- Picciotto, V. (2013). *Consciousness and mental quotation: An intrinsic higher-order approach*. Ph.D. thesis.
- Putnam, H. (1988). *Representation and reality*. Cambridge, MA: MIT Press.
- Quine, W. V. O. (1976). *The ways of paradox, and other essays*. Harvard: Harvard University Press.
- Rosenthal, D. (2005). *Consciousness and mind*. Oxford: Clarendon Press.
- Russell, B. (1927). *The analysis of matter*. London: Routledge.
- Schmidhuber, J. (1993a). A 'self-referential' weight matrix. In: *International conference on artificial neural networks* (pp. 446–450). Springer.
- Schmidhuber, J. (1993b). An 'introspective' network that can learn to run its own weight change algorithm. In: *1993 third international conference on artificial neural networks* (pp 191–194). IET.
- Seager, W. (2006). The 'intrinsic nature' argument for Panpsychism. *Journal of Consciousness Studies*, 13(10–11), 129–145.
- Seager, W. (1995). Consciousness, information and Panpsychism. *Journal of Consciousness Studies*, 2(3), 272–288.
- Seager, W. (2016). *Theories of consciousness: An introduction* (2nd ed.). Milton Park: Routledge.
- Searle, J. R. (1992). *The rediscovery of the mind*. Cambridge: MIT Press.
- Shoemaker, S. (1982). The inverted spectrum. *The Journal of Philosophy*, 79(7), 357–381.
- Strawson, G. (2019). What does "physical" mean? A prolegomenon to physicalist Panpsychism. In: W. Seager (Ed.), *The Routledge handbook of Panpsychism*. Abingdon: Routledge.
- Szangolies, J. (2018). Von Neumann minds: A toy model of meaning in a natural world. In: A. Aguirre A, B. Foster, Z. Merali (Eds.), *Wandering towards a goal* (pp. 29–39). Springer.
- Szangolies, J. (2020). The abstraction/representation account of computation and subjective experience. *Minds and Machines*, pp. 1–41.
- Szangolies, J. (2015). Von Neumann minds: Intentional automata. *Mind Matter*, 13(2), 169–191.
- Turing, A. M. (1937). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 2(1), 230–265.
- von Neumann, J. (1966). The theory of automata: Construction, reproduction, homogeneity, original unpublished manuscript, 1952–1953. In: Burke AW (ed.), *Theory of self-reproducing automata* (pp. 91–381). University of Illinois Press, Urbana, IL.
- Waters, D. P. (2012). Von Neumann's theory of self-reproducing automata: A useful Framework for Biosemiotics? *Biosemiotics*, 5(1), 5–15.
- Whyte, J. T. (1990). Success semantics. *Analysis*, 50(3), 149–157.
- Williford, K. (2006). The self-representational structure of consciousness. In U. Kriegel & K. Williford (Eds.), *Self-representational approaches to consciousness* (pp. 111–142). Cambridge, MA: The MIT Press.
- Yudkowsky, E., & Herreshoff, M. (2013). Tiling agents for self-modifying AI, and the Löbian obstacle. Early draft. Technical report. Machine Intelligence Research Institute, Berkeley, CA, <http://intelligence.org/files/TilingAgentsDraft.pdf>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.