Entropy **2009**, *11*, 295-325; doi:10.3390/e11020295

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

A Universal Framework for Analysis of Self-Replication Phenomena

Bryant Adams[1,2,*] **and Hod Lipson**[3]

[1] Department of Mathematics, Cornell University, Ithaca, NY, USA

[2] Department of Mathematical and Physical Sciences, Wells College, Aurora, NY, USA

[3] Departments of Mechanical & Aerospace Engineering, and Computing & Information Science
Cornell University, Ithaca, NY, USA; E-mail: hod.lipson@cornell.edu

[*] Author to whom correspondence should be addressed; E-mail: badams@wells.edu

Received: 24 March 2009 / Accepted: 19 June 2009 / Published: 22 June 2009

Abstract: In this paper, we propose definitions for a general, domain-independent concept of *replicability* and specifically focus on the notion of self-replication. We argue that self-replication should not be viewed as a binary property of a system, but rather as a continuously valued property of the interaction between a system and its environment. This property, *self-replicability*, represents the effect of the presence of a system upon the future presence of similar systems. We demonstrate both analytical and computational analysis of self-replicability for four distinct systems involving both discrete and continuous, formal and physical behaviors.

Keywords: self-replication; self-reproduction; automata

1. Overview and History

Self-replication is a fundamental property of many interesting physical and formal systems, such as crystals, waves, automata, and life [1]. Despite its importance to many phenomena, self-replication has not been consistently defined or quantified in a rigorous, universal way. Two prominent issues arise in examining how self-replication has been handled when trying to extend the concept universally: how to deal with non-ideal systems and how to address so-called ‘trivial’ cases [2]. Here we propose that both these difficulties can be overcome by treating self-replication as a continuous property [3].

First, when making the transition from idealized or formal systems to concrete ones, we face systems that are too complex to model or whose models still would lack convenient simplifying elements. For example, Moore[4] requires that for a cellular automata configuration to be considered self-reproducing, it must be capable of causing arbitrarily many offspring. Given an abstract or idealized context, this does capture part of the long-term nature tied up in an intuitive conception of replication; however, it extends poorly to finite (even if very large) contexts or natural ones that might involve factors such as predation. Lohn and Reggia [5] put forward several cellular-automata-specific definitions for self replication, such as distinct copies being initially adjacent but eventually separable, and having unbounded growth potential, but these concepts are again difficult to extend to non-formal systems, and only allow for a binary labeling.

A second issue that arose in the consideration of self-replicating automata was that some cases seemed too trivial for consideration. This triviality is derived from the burden of replication being placed on the context, rather than being shared in some nontrivial way by the self-replicating entity itself. An example of such a system is a cellular automaton whose rule activates all cells that had at least one active neighbor. A system defined as 'an active cell' could certainly give rise to an arbitrarily large number of new active cells, but such a case does not seem to exemplify what we seek to capture in the concept of significant self-replication.

In compiling Von Neumann's work, Burks [6] sought to address this second concern. Burks stated that "Clearly, what is needed is a requirement that the self-reproducing automaton have some minimal complexity." As a bound on the complexity, he used the requirement that a "self-reproducing automaton also be a Turing machine." As Langton [7] points out, however, this not only "eliminates the trivial cases, but it also... eliminates all naturally occurring self-reproducing systems as well," because none have been shown equivalent to a Turing machine. While claims to the contrary do exist [8], this does suggest that a view not reliant on a requirement for universal computation could be worthwhile.

A third issue, not as prominent as the first two but useful in motivating some of our choices, is the apparent chicken-and-egg situation arising when, as Löfgren puts it, "we ask not merely how a cell can reproduce in a suitable surrounding, but how this property has evolved" [9]. When simply the inputs and outputs of self-replicating systems are considered (as in a brief robot example on page 299) and the systems themselves are represented as mapping functions between these spaces, one finds that a 'complete' self-replicating system is represented by a function which is a member of its own range. In 1959, Rosen argued [10] that this produces a paradoxical situation: the function (replicating system) cannot exist until its range (which includes the system) is fully defined, but the range (which must include the system) cannot be defined until the function exists.

As introduced by Rosen, the distinguishing details present in the physical world of distinct physical structures and distinct points in time are not considered (making Löfgren's 1968 resolution to the paradox a stronger result) but in 1966, Guttman raised an alternate route to resolving the paradox - including additional information present in real-world scenarios. By not being required to reproduce 'identical' (i.e. non-identical at more than a physical and temporal level, but being distinguishable at a some unspecified, grosser level) offspring, Guttman essentially argued that chickens and eggs could co-evolve. While his particular presentation in [11] had deficiencies raised by Löfgren, we follow Guttman's example by counting replication results which are 'close enough', and by distinguishing between a system at a given time and an 'identical copy' of the system at a later time.

Combined, criteria arising from these issues led to a coarse-grained (binary) definition of self-replication [12]. The definition for self-replicability we propose here is motivated in part by:

- A desire to capture self-replication as more than just a binary property which is applicable only to certain automata;

- The goal of encapsulating a general concept in a means not reliant upon (but compatible with) ideal conditions.

We wish to do this by putting self-replication on a scale that is algorithmically calculable, quantifiable, and continuous. Such a scale would allow for comparisons, both between the same system in different contexts, determining ideal contexts for a system's replication, as well as between different systems in the same context, if optimizing self-replicability in a given context is desired.

Rather than viewing self-replicability as a property purely of the system in question, we view it as a property of the interaction between a system and its context. This does raise the concern of allowing what might be considered trivial replicators, such as a factory which automatically replicates anything that is placed in it. The low self-replicability found in the second (ring system) example, however, gives hope that 'trivial' cases may distinguish themselves by small self-replicability that behaves like $\log \frac{c+t}{t}$ as time t becomes large, where c is a constant.

Self-replication – as we present it – is a property embedded and based upon information, rather than a specific material framework. In physical systems, the replicas are only similar to the original material pattern but use distinct material components. In the case of a replicating program, the memory locations used to embed the information are distinct. Many of the most interesting systems, such as living organisms, create significantly different offspring, despite passing along very similar genetic information.

We also take the view of Sanchez *et al* [13] in seeing a distinction between replication and reproduction. Specifically, replication seeks to copy an entire system without error, while reproduction allows for variations. Still, systems that self-reproduce do something akin to self-replicating, especially when the concept of 'self' is interpreted more loosely.

We will construct replicability as a property relative to two different contexts, which indicates the degree to which one context yields a higher presence of the system over time. *Self*-replicability, then, is a comparison between a context lacking the system and a context in which the system is present. After a discussion of the terms "self-replication" and "self-reproduction", we will introduce a number of definitions, and then give examples of determining the self-replicability of four types of systems:

- First, in section 4.1., a strictly bounded, formal, finite system, which can easily be calculated explicitly.

- Second, in section 4.2., a model of a physical system with discrete time but continuous location. This example is also explicitly calculated, with values derived from computer simulation.

- Third, in section 4.3., a model of a continuous, probabilistic, physical system, wherein the calculations are based on analytical abstractions of the behavior of the system, without actually reproducing its behavior.

- Fourth, in section 4.4., a model of L.S. Penrose's self-reproducing machines. In this case, the behavior of the system is reproduced in the model and measurements are made on the concrete (modeled) population rather than an abstraction.

2. Conceptual Framework

2.1. Reproduction vs Replication

Intuitively speaking, living systems self-reproduce. Intuitively speaking, artificial systems (primarily in the form of computer code but with growing momentum in the physical domain) self-replicate. Machines can, at least theoretically, make perfect replicas of themselves, and in order to preserve the delicate nuances of their inner workings *must* make those replicas identical to within some collection of exacting tolerances. Biological life does, in a broad range of observable examples, make similar-but-distinct offspring which share a common overall genetic code and overall morphology, but which can be easily differentiated from the parent on many scales.

Still, when a robot replicates itself, it might be said to be producing something already produced. Namely, itself. On the other hand, when a baby owl hatches there is some underlying genetic information that captures owl-ness that has been replicated. While self-reproduction usually implies mutability and evolvability, an excess of change from one generation to the next, i.e. a lack of sufficient replication, can go from self-reproduction to simply spawning nonviable mutant cells. While self-replication usually implies invariance, at some level there are always differences, even if it requires pointing out that a computer virus is being stored in different physical locations, or the molecules composing a split-off bacterium are not all the same as those found in the parent.

Where do self-replication and self-reproduction stand, relative to each other? We'd like to argue that self-replication, in a strict sense, should be considered a nigh-unreachable limiting case of increasingly variance-free reproduction. And, as such, that reproduction may be viewed as more or less inaccurate replication. Further, that the natural domain in which self-replication should be considered is that of underlying information, rather than particular physical instantiations.

Looked at in the strictest sense of creating completely identical but still separate offspring, there can be no such thing as self-replication. Reducing the argument to the absurd, the Pauli exclusion principle leads us to physically distinct realizations of our items of interest, and that which can be distinguished is not completely identical. Slightly less absurd, one might require that replicants be identical at the quantum level to be considered truly the same, but Pati and Braunstein show that finite resources are insufficient for a perfect quantum-level universal constructor[14]. Clearly, the bar is being set too high when requiring completely identical copies.

So, how far is it appropriate to lower the bar? Perhaps the molecular configuration must be identical, but the molecules and their positions don't have to be the same. Bacteria with any floating organelles will probably be almost included, but not quite. Viruses could be included in self-replicators, if they didn't mutate so much. Machines, however, would almost certainly be ruled out, barring those constructed one molecule at a time. Computer viruses are hard to even fit into this conception, as they are characterized better by abstract information than physical configuration. It seems that the qualifications for 'replicant' is still too high to include the cases that ought to count. Moreover, having a qualification based on physical properties rules out discussing abstract systems at all.

Lower the bar too much, though... enough to allow for machines that match a blueprint within given tolerances, for example... and we risk stepping into the land of reproduction. Parthenogenesis, for example, is an asexual form of reproduction that can result in genetically identical offspring, which certainly falls under the heading of "matching a blueprint within any given tolerance", so should we say that parthenogens self-replicate? It seems difficult, just looking in the physical domain, to draw a clear and consistent line between replication and reproduction. And, even if that weren't the case, the placement of less tangible replicators/reproducers such as computer virii or social 'memes' complicates the matter further.

As we take the view in this paper that reproduction and replication are not solely properties of physical constructs, we feel it is important to include intangible cases in our consideration. In doing this, we are lead to see replication/reproduction more as a property of information (machine blueprint, genetic code, memetic structure) than the particular expression of that information. In some ways, this makes semantic life a little less confusing, and allows for the return to a strict notion of replication.

For example, humans reproduce. Human offspring are (obviously) not identical to the parent, and so do not fit into the strict domain of replication. The reproduction process does involve the copying of genetic material but, even if no mutations occurred, the copies would still be distinguishable based on their physical location. The digital information carried by those copies, however, is replicated: after reproduction, there are now additional physical configurations (nucleic acids) which adhere to the same abstract patterns (the order of nucleotides determining a particular allele). At the level of unmutated genes, we see strict replication. At larger levels, of course, the genetic information differs from that found in either parent, and is neither being strictly replicated nor colloquially reproduced. By specifying a broader abstract pattern (i.e. collections of genetic information which code for a human) however, and not requiring identical results but only "ϵ-close" results which fall into that specified range, we may widen our view to allow for ϵ-replication. Whether the particular mechanism by which there came to be more genetic material coding for a human was through sexual reproduction that involved a combination of gene replication and mutations, or through a deliberate construction in a lab of the desired DNA sequences, there is now more genetic material whose information lies within the "ϵ-ball" of patterns which code for humans.

As this example indicates, our approach does not seek to directly take into account the particular biological mechanisms by which reproduction occurs, when considering the considering the replicability of a living organism, nor would it be concerned with the particular steps by which a robot self-replicated (or whether the robot itself or its environment were 'really' performing the work of replication). Rather, it will be concerned with how a population's size over time is dependent upon members of the population being present in the first place. Explicit understanding of the complex and case-specific processes involved may be necessary to perform our analysis in some cases (specifically, to construct an appropriate time development function (Definition 3)), but in general we look at both replication and reproduction on the level of "getting more of similar things."

To take a more contrived example, three hypothetical robot types (A,B,C) are designed with two abilities. Each robot has some (possibly empty) *construction set* of other robots it is capable of directly constructing, and each robot has some (possibly empty) *modification set* of robot types that can be changed to other robot types. Let us suppose that type A has construction set $\{B\}$ (Type A robots can produce type B robots) and modification set $\{(B \mapsto C)\}$ (Type A robots can turn type B robots into type C robots), type B has construction and modification sets $\{\}$ (unable to make or alter anything), and type C has construction set $\{\}$ and modification set $\{(B \mapsto A), (A \mapsto B)\}$ (can change type A to type B and vice versa).

In this case, starting with a type A robot, another type A robot can be produced: The type A robot produces two type B robots, then modifies one of them into a type C robot, which in turn modifies the other type B into a type A. At no stage, however does a robot directly bring a replica of the same type into being. At best, a type A robot 'reproduced' a couple B robots (though this requires a very inclusive notion of reproduction), one of which was acted upon in such a way that it 'grew up' into another type A robot. Yet, despite all the variations and indirectness, the end result is that there are more robots which match the 'type A' pattern. Viewed through the lens of "robot A nontrivially self-replicates if, without significant work on the part of the environment (such as a C robot) it can build another A robot", it may be unclear whether such mediated robot production counts as self-replication or not. When we focus instead on the presence of specified patterns, however, this is clearly another case of self-replication, regardless of which robots are considered the active agent.

Continuing with the emphasis on underlying information, consider the gradual notion of speciation. Supposing, for example, that horses and donkeys had common ancestry: There was a point when pre-horses and pre-donkeys were still mutually viable subspecies. At that point, some pre-horse offspring would have also qualified as pre-donkeys, and some would be discernibly not pre-donkey material. Similarly for pre-donkeys, it would have been a tough call in some cases to say if a pre-donkey, a pre-horse, or both had been reproduced. Considered as design spaces, however, an overlap is not such a problem. Some offspring will be expressible only as representatives of a single design space, some will be expressible within either design space. A pre-horse replicates the horse design space, and it is nothing worse than an interesting coincidence that it happens to also be in the donkey design space.

Overall, we will look at self-reproduction as more or less imperfect self-replication, or at self-replication as a limiting case of arbitrarily similar reproduction. Moreover, we tend to take more interest in the underlying abstract information which may be undergoing replication, than in the physical manifestation which, relative to its 'parent' physical manifestations, is undergoing reproduction.

2.2. What is being Replicated?

With placement of reproduction/replication in mind, our conception of self-replication is in large part concerned with the information and patterns within a system as the core item of interest, rather than particular physical expressions. Highlighting this view is part of the motivation for two of our examples, namely the second (patterns in wave propagation) and third (crystal growth) models. In general terms, note that while both crystals and infants grow, there is something markedly different about their forms of growing. A baby cannot be clearly seen as a number of smaller babies pieced together, or as an even larger number of mutually overlapping babies. A crystal, on the other hand, may bear structural self-similarities on multiple scales, and while (unless broken) there is still 'one crystal' growing, the amount of crystal-*pattern* present increases much more as a crystal grows than the amount of baby-pattern increases as a child ages. So, while the size of a crystal is merely growing, the underlying crystal-pattern is undergoing replication.

Given that the growth of a baby coincides with an increase in the presence of its genetic material, it may be tempting to say that the situations are not actually distinct. The ground for comparison, however, is "As a X grows, how much does the amount of X*-pattern* present increases?" We must make uniform replacements for X, and cannot selectively replace the first X with "baby" and then the second X with "genetic material", unless we consider them to be one and the same thing. We can compare between when X =baby and when X =crystal, and doing so yields different answers. This highlights that it is important to maintain a clear view of our object of interest, whose replicability is being considered. Toward this end, a 'presence' function will be introduced to measure the quantity of our object that is found in a given context, and this will rely on an assumed ability to measure dissimilarity.

Exactly what dissimilarities are and are not taken into consideration is a choice that must be made, and this choice could vary greatly depending on intent. If we do want to consider a baby to represent the same pattern as its genetic material, dissimilarity would be measured differently than if we view them as distinct. If we are interested in physical replication resulting in spatially separate offspring, dissimilarity would be measured differently than if we were interested in the replication of a certain pattern across arbitrary scales.

2.3. *Identifying Triviality and the Importance of Context*

A pervasive issue with discussing self-replication is if a definition includes trivial self-replicators. When dividing the universe into things which either do or do not self-replicate, there is little room for this troublesome middle ground. Our conception of self-replication does not directly recognize any line distinguishing 'trivial' from 'non-trivial' replicators, answering the question of "Is X a trivial replicator?" with "It depends on the context."

The context in which a system's self-replicability is measured is of major importance in our conception of the term. Intuitively, snowshoe hares are good (albeit, as natural systems, error-prone) replicators in the context of a northern forest, they are poor replicators in a desert which they are ill-equipped to survive, and they are abysmal replicators in deep space. It is not enough to simply say they replicate, or even that they replicate well, because these statements only hold in certain contexts.

A usual identifier for a trivial replicator is that it does no work on its own, but relies on some external feature to 'do all the work'. A computer program which directly calls the operating system's 'copy' routines might be considered trivial. A single activated cell in a cellular automata whose rule stipulated that "If a cell or any of its neighbors are active at time T, the cell is active at time $T + 1$" might be considered trivial. A biological virus which carries genetic instructions but no machinery or resources with which to follow those instructions might be considered trivial.

Yet, at a lower level, cells and viruses do just as much of their own work: they sit back and let physics run its course and happen to both be configurations which, in the course of physical laws unfolding, cause replicas to appear. Every self-replicating pattern in a cellular automata takes advantage of the CA's rules, which is what's doing all the work. Classifying triviality on the basis of "taking advantage of the rules of the context" is far too broad.

Visibly, just because a system makes extensive use of its context, it does not necessarily deserve to be called trivial. To take a recent example, Chirikjian, Zhou, and Suthakorn's 3 module, 4 component replicating Lego robot [15] should have a large self-replicability when put in the very specific context that includes parts in the right locations, accompanied by a track to run along. In any other context, though, its self-replicability should be just about zero. A 'smarter' and more sensor-laden robot which could find and assemble pieces without the use of tracks to follow would be *less* trivial, as it would use even less context-based artifacts, just as a 'dumber' robot which only ran along pre-laid tracks which automatically switched properly would be more trivial.

Or, taking viruses as a conceptual model, a context-sensitive measure of self-replication could distinguish more or less 'trivial' systems in the following sense: consider what 'minimal' contexts are necessary for each of the replicators to achieve a certain level of self-replicability. For a cell, this would include some raw materials and energy, while for a virus this would include some raw materials, energy, *and* a cell with access to those resources in which the virus could operate. Given that a virus requires a more complex context, i.e. one with more built-in 'machinery', a virus would be considered 'more trivial'.

2.4. Comparison to Another Metric

Other tools have been developed for measuring the complexity, information content, structural organization, and degree of self-replication of various systems. Though directed at automated design rather than self-replicating systems, Hornby compares[16] a number of metrics which could have correlations with measures of self-replication. In order to make comparisons between different self-replicating robots, Lee and Chirikjian define[17] a degree of self-replication in terms of the complexity of active elements and interconnections within a system. Many choices come up in defining a framework, and a comparison between the approach taken in this paper and Chirikjian's measures reveals a number of different approaches:

Information about design: The measures in this paper are extrinsic, in the sense that while they will require a distinction between what is and what is not the system in question, they do not depend on knowledge about how it works. Chirikjian's measure is intrinsic, in the sense that it explicitly makes use of information about how systems are constructed and internally arranged.

Specificity: The measures in this paper are aimed at systems which self-replicate in general, while Chirijkian's measure is very well suited for application to modular robotic systems but not clearly extensible to self-replicating systems such as prions or memes whose components are less well defined.

Grounding: The measures in this paper are based on differences in population, where Chirikjian's measures focus on differences in complexity. Insofar as the consolidation of resources into a particular population of replicators is correlated with a change in entropy, the measures may share common ground. In general, Chirikjian's measure is more appropriate for analysis of an individual replicator, while this paper's notion of self-replicability is geared toward replication-fitness within a population.

Role of 'self': Chirikjian et al. say[17] "If an external agent partially or fully controls the replication process, we call the system replicating" but not "self-replicating"." In this paper, we intentionally abandon concerns with the source of control, and say that a system is "self-replicating" when "the current presence of a system positively influences the replication process."

Role of environment: Similarly, in a case such as Zykov et al.'s self-replicating 'molecube' robots[18], the measure in this paper does not distinguish between an environment with an arbitrarily large supply of a resource available from a particular location and the simulation thereof by a human placing resources as they are consumed, while it appears that Chirikjian's measure would not classify the latter as self-replication, due to the presence of an external agent, but would classify the former as self-replication.

3. Definitions

The definitions that follow, while they lead to discussion of self-replicability *of a system* for this paper, are designed with the general consideration of relative replicability *of a system in different contexts* and comparison between *different systems in a particular context* in mind. Our focus here is on the specific case we wish to call self-replicability, but want the larger 'relative' concept to remain visible in the periphery. As such, not all the definitions presented will be directly used: While a number of intermediate steps could be compressed into a longer explanation of Definition 12 (Self-replicability), we feel that the smaller steps make the intended structure easier to grasp, as well as revealing directions in which other aspects of replicability could be investigated.

We wish to capture the notion that "Self replicability is a measure of how influential a system is, within a given context, in establishing a larger future population of sufficiently similar systems." To construct such a measurement, we first need to get a handle on the notions of 'a context', how to determine if systems are 'sufficiently similar', and how to compare population sizes in a general sense which does not rely on countable, distinguishable individuals.

Definition 1 (Context)**.** By *Context* we denote a single state of a (presumably closed) system.

For example, with a configurationally closed system such as a binary 5×5 grid, each of whose cells may be either 'on' or 'off', one context, E_1, would be 'all 25 cells off', while another, E_2, might be 'every other cell on'. We note before continuing that we will be regarding 'systems' whose replicability is of interest as subsystems of a context, rather than forming a system-context dichotomy.

For physical systems, a context is essentially a snapshot of the physical environment in which the system is sitting, hence the use of 'E' for contexts. While it could be called 'the state of the system(environment) in which the system(putative replicator) is being examined.', we introduce the term *context* to improve clarity and for ongoing emphasis on the role it plays in discussing replication.

Definition 2 (Set of configurations)**.** Given a context E of some system, the *set of configurations* $\overline{E}$ is the set of all possible states of that system. We call elements of $\overline{E}$ "*E-configurations*".

Using the 5×5 grid example, the set of configurations would be the set of all 2^{25} states the grid could assume. Note that, if for some contexts E_1, E_2, we have $E_1 \in \overline{E_2}$, then it also holds that $E_2 \in \overline{E_1}$ and $\overline{E_1} = \overline{E_2}$.

Definition 3 (Time development function)**.** A *time development function* is a map $T : \overline{E} \times \mathbb{R}^+ \rightarrow \overline{E}$, (respectively $T : \overline{E} \times \mathbb{Z}^+ \rightarrow \overline{E}$ for discrete time systems) constrained by $T(E, 0) = E$ and $T(T(E, y), x) = T(E, x + y)$.

With a time development function, we operate under the assumption that the progression between states as time passes is externally deterministic. We also assume there is no difference between a system that 'ages' five units and then ten units, and a system starting in the same state that first ages ten units, and then ages five units. When we write only $T(E)$ rather than $T(E, n)$, it is assumed there is a natural increment of time and we are using $T(E, 1)$.

Note that the set $\{T(E, x)|x > 0, x \in \mathbb{R}\}$ is the set of all contexts which E could become, while $\{Y|\exists x \in \mathbb{R} \text{ s.t. } T(Y, x) = E\}$ is the set of all contexts which E could have arisen from.

Definition 4 (Subsystem set)**.** If S is a system, we denote by S^* the set of all subsystems of S, and say S^* is the *subsystem set* of S. We extend this notation to say that if X is a state of S, then we denote by X^* the set of states $\{X' \in S' : S' \in S^* \; and \; X|_{S'} = X'\}$. That is, all states of subsystems of S which are the same as X wherever that comparison is meaningful.

Often, we are most interested in the X^* where X is a context. For example, in the case where the system S is a 5×5 grid whose points can be either 'on' or 'off', and E is a state with all points being off, E^* would include, among its 2^{25} elements, four 4×4 binary grids with all points being off, five 1×5 binary grids with all points being off, and a number of L-shaped binary grids with all points being off. For further example, given a second system E_2, a 5×5 grid in which the top row of points were on, the rest off, E_2^* would still have 2^{25} elements, each the same shape as a corresponding element in E^*, but now some elements would include 'on' points if they included any of the top row.

Definition 5 (Possible Subsystems)**.** We define the *possible subsystems* $\overline{E}^*$ as the union of all F* such that $F \in \overline{E}$.

In the case of the binary grid, the elements of $\overline{E}^*$ could be classified into 2^{25} distinct shapes, and each shape-class would have 2^n elements, with n being the number of cells in that shape.

Definition 6 (Dissimilarity pseudo-metric)**.** To quantify the 'self' portion of self-replication, we assume that a *dissimilarity* pseudo-metric $d : \overline{E}^* \times \overline{E}^* \rightarrow \mathbb{R}^+$ is given. Recall a pseudo-metric d obeys $d(x, y) + d(y, z) \geq d(x, z), d(x, y) \geq 0$, $d(x, y) = d(y, x)$, and $d(x, x) = 0$, but not $d(x, y) = 0 \implies x = y$.

Note that d induces an equivalence relation on $\overline{E}^*$. That is, for $S_1, S_2 \in \overline{E}^*$ we say that $S_1 \equiv S_2$ exactly when $d(S_1, S_2) = 0$. Presumably, the dissimilarity pseudo-metric would be chosen such that it induced a natural equivalence relation, but this choice is not assumed.

Definition 7 (Presence)**.** We define the *presence* $P_\epsilon(E, S)$ *of a subsystem* S *in a context* E *within tolerance* ϵ to be the probability that a randomly selected subsystem $T \in E^*$ will satisfy $d(T, S) \leq \epsilon$. Formally, if μ is a probability measure with $\mu(E^*) = 1$, $\mu(\emptyset) = 0$, and $\mu(U) \in [0, 1]$ for U a measurable subset of E^*, then $P_\epsilon(E, S) = \mu(\{T \in E^* : d(T, S) \leq \epsilon\})$

Essentially, the presence function measures 'how much' S is found in E. As a probability, P takes values in the interval $[0, 1]$. In the case of discrete contexts, the presence function essentially reduces to counting how many copies of S (for zero-tolerance) or S-like things (for nonzero tolerance) can be found in E (and then dividing by the total number of things in E.) In a continuous case, however, even a countable infinity is still negligible, and thus we (temporarily) increase the measure of the set by using a nonzero tolerance.

Definition 8 (ϵ-present, ϵ-possible)**.** When $P_\epsilon(E, S) \neq 0$, we say that S is *ϵ-present* (in E). Also, we say that S is *ϵ-possible* (in E) when there is some time $t \in \mathbb{R}^+$ such that S is ϵ-present in $T(E, t)$.

Definition 9 ((Relative) Replicability, Momentary)**.** Given a set of configurations $\overline{E}$ and two E-configurations E_1, E_2, we define the *momentary relative replicability* of a system S in E_1 relative to E_2 with tolerance ϵ at time t as

$$R_M(S, E_1, E_2, \epsilon, t) = \log \frac{P_\epsilon(T(E_1, t), S)}{P_\epsilon(T(E_2, t), S)} \tag{1}$$

The ratio in Eq. (1) serves to compare the probability at time t of finding S in the future of E_1 to the probability at the same time of finding S in the future of E_2. There are a few cases where Eq (1) is undefined. If $P_\epsilon(T(E_1, t), S) = P_\epsilon(T(E_2, t), S)$ (including zero) we define R_M as 0: systems are equally present when they are both completely absent. When $P_\epsilon(T(E_1, t), S) = 0$ but S is ϵ-present in $T(E_2, t)$, we define R_M as $-\infty$, indicating that the presence in E_2 is greater than E_1 in a trivial sense, and when $P_\epsilon(T(E_2, t), S) = 0$ but S is ϵ-present in $T(E_1, t)$, we similarly define R_M as ∞.

In the case where S is not ϵ-possible in either or both of E_1 and E_2, we will not define replicability. It is not meaningful to discuss how well a system does anything, in environment in which that system cannot possibly exist.

Our choice of a logarithmic scale is not simply aesthetic: while some self-replicating systems might have a linear population growth, many are polynomial or exponential. Without a scaling method which can reduce exponential growth to linear growth, comparisons, as time became large, could yield arbitrarily large differences due to as little a change as the population starting its growth at a slightly later time. Thus, to allow meaningful comparison between potentially exponential growth rates, we view every-thing in a logarithmic scale.

Note that, given a system whose population is not always 'present' (adult mayflies, for instance), the Momentary Replicability could fluctuate wildly, even though over long time spans, the behavior of the population may be well behaved. We wish to have an overall, long-term picture of our system's influence, but simply averaging momentary replicabilities would give excessive weight to fluctuations that are essentially artifacts of our sampling locations. Thus, rather than extend from momentary to long-term by directly combining multiple momentary comparisons, we combine our population measurements over a whole range of time first, then make a single comparison:

Definition 10 ((Relative) Replicability, Over Time)**.** In the case when S is ϵ-possible in both E_1 and E_2, we also define the *replicability* over time τ_0 to τ_1 (in E_1 relative to E_2 with tolerance ϵ) as:

$$R_T(S, E_1, E_2, \epsilon, \tau_0, \tau_1) = \log \left(\frac{\int_{t=\tau_0}^{\tau_1} P_\epsilon(T(E_1, t), S) dt}{\int_{t=\tau_0}^{\tau_1} P_\epsilon(T(E_2, t), S) dt} \right) \tag{2}$$

Note that for discrete systems, the integrals reduce to sums.

This is our most general and flexible notion of replicability, allowing considerations of various durations of interest, comparison between the system's performance in two different contexts, and at an unspecified tolerance for error. To move toward our notion of Self-replicability at this point, we first need to discard dependence upon a particular choice of time. Since our target idea involves the influence of a system upon it's future presence in general, the natural first restriction is to consider all times in the future:

Definition 11 (Replicability, Overall)**.** We define the *Overall Replicability* as the limiting case:

$$R_O(S, E_1, E_2, \epsilon) = \lim_{t \to \infty} R_T(S, E_1, E_2, \epsilon, 0, t) \tag{3}$$

Note that, by using the logarithm of the fractions, we have an additive form for some basic relations. Letting $R(S, E_1, E_2)$ stand in for any of the defined types of replicability (exercising due caution when R takes on infinite values), letting S be a fixed system, and letting $A, B, C \in \overline{C}$, we have:

$$\begin{aligned} R(S, A, B) + R(S, B, C) &= R(S, A, C) \\ R(S, A, B) &= -R(S, B, A) \\ R(S, A, A) &= 0 \end{aligned} \tag{4}$$

While trivial to derive, the last relation highlights that replicability is being taken as a relative, rather than absolute, concept. When a replicability value is zero, we have two contexts in which a system fares equally well. As the replicability increases, we see the system faring better in the first context.

In order to approach the idea of *self*-replicability, we will specifically consider cases where S is present in E_1, is not present in E_2, and $d(E_1, E_2)$ is minimal (among the choices satisfying the presence conditions). In these cases, we are making a comparison between a 'blank' context and one that is minimally different, in order to reflect a minimally disturbing introduction of a system into the context. For example, in modeling a crystal, we might consider a supersaturated solution as E_2, and a supersaturated solution with a tiny seed crystal as E_1, but we would not consider a fully crystallized configuration for E_1.

In particular, let a context E and a system S be given, such that S is minimally ϵ-present in E. That is to say there are no contexts similar to E in which S is strictly less ϵ-present or, formally, there exists a δ sufficiently small that for all E' with $d(E, E') < \delta$ we have $P_\epsilon(E, S) \leq P_\epsilon(E', S)$. Let $E' = E - S$ denote some $E' \in \overline{E}$ such that S is not ϵ-present in E' but is ϵ-possible in E', and $d(E, E')$ is minimal. This leads to the particular case of self-replicability we wish to examine.

Definition 12 (Self-replicability, Overall)**.** With E, S, and $E - S$ given as above, we define the *self-replicability* of a system S in a context E as

$$R_S(S, E, \epsilon) = R_O(S, E, E - S, \epsilon) \tag{5}$$

Note that, when $E - S$ cannot be uniquely defined, R_S is not uniquely defined. However, this concern does not arise until we leave the safety of reasonably homogeneous contexts.

In essence, R_S looks at how present S becomes after it was at one point specifically present in the context, as compared to how it develops without any prompting.

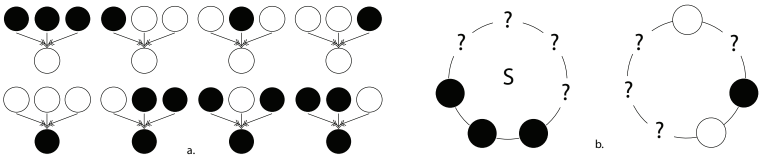
To ground Overall Self-replicability conceptually, we return to the notion we wish to capture: "Self replicability is a measure of how influential a system is, within a given context, in establishing a larger future population of sufficiently similar systems." If R_S is zero, then the inclusion of S neither adds nor detracts from the future presence of S. As such, the presence within a given context of a system which has zero self-replicability within that context has no influence on the future population of similar systems. If R_S is positive, then its presence has a (positive) effect upon future populations of similar systems. Conversely, a negative value indicates that including the system has a detrimental effect, and that the system is essentially self-defeating. Infinite values, both positive and negative, will arise in cases where the lifespan of a system is finite in one of E or $E - S$ and infinite in the other. Here, we say a system S in a context E has a *finite lifespan* when there is some time t such that S is not ϵ-possible in $T(E, t)$. In particular, when a system could not possibly arise in some context through the natural time-development of that context, but can only arise through having been artificially inserted, overall replicability will be infinite.

4. Examples

We now consider four example cases. The first is a two-state cellular automaton that can be explicitly simulated and calculated. Second, we will consider a simulated optical fiber ring that mixes discrete and continuous elements. Third, we look at an abstracted model of an expanding system, demonstrating a case of calculation without explicit simulation. Finally, we present an analysis of L.S. Penrose's self-reproducing machines.

4.1. Cellular Automaton

Figure 1. (a) Table of the 8 evolution rules defining a one dimensional CA of radius 1 and (b) Two examples of subsystems. We will calculate the self-replicability of the one labeled S.



Our first system is a cellular automaton given by single-cycle graph with seven nodes, each of which can take the state of filled or hollow, along by a radius-1 evolution rule (shown in Fig. 1a). Up to rotations and reflections of the graph, there are eighteen distinct states the system can take. There are three limit cycles under the evolution rule: First, the 'all filled' state is sent to the 'all hollow' state, which is then sent to the 'all filled' state again. Second, there is a two-cycle which alternates between two configurations having no symmetry (Fig. 3). Finally, there is a large fourteen-cycle (Fig. 2).

Examples of two subsystems are shown in Fig. 1b. As this is a discrete system, we can move directly to the zero-tolerance case, and thus need to define only the preimage of zero under the dissimilarity pseudo-metric (i.e. establish the meaning of 'identical'.) There are 128 configurations, of which only eighteen are distinct, and each configuration has 128 subsystems. We define a dissimilarity pseudo-metric that is zero when the subsystems being compared have, up to rotation and flipping, the same shape and point-wise parity. Note in fig. 1b, the left subsystem only requires that three adjacent nodes be black in order to match, the remaining nodes can be in any configuration. Likewise, the right subsystem only 'looks at' three nodes at a time.

The time development function is created by defining $T(E, 1)$ as the result of the automata evolution function being applied once to E, and then extending inductively with $T(E, n) = T(T(E, n-1), 1)$. In this discrete case, the presence function simply counts the number of matches and divides by the total number of subsystems, and the divisor is constant for all calculations of presence. (This will not be the case in the unbounded example later.) Consequently, we will take advantage of a later cancellation and count the matches in place of using the presence.

Thus equipped, we will calculate the self-replicability of the system S consisting of three adjacent cells which are all black, in the context E_1 with four adjacent points filled and the rest hollow, with respect to E_0, a context with one of the middle two points set as hollow. First, we note that S is possible in E_0, since it is present in $T(E_0, 1)$, and S is clearly possible in E_1, since it is present in E_1. Also, note that while we have not defined the dissimilarity pseudo-metric for non-identical cases, one would expect that, having fixed E_0, a natural pseudo-metric would yield E_1 as a minimally dissimilar context among those containing S, as it differs in only one location. Note that the choice of removal here is important: working through the following steps when one of the filled cells on the end of a four-cell block is set hollow rather than a middle cell, will yield a zero replicability!

Figure 2. Successive states of E_1 under T.

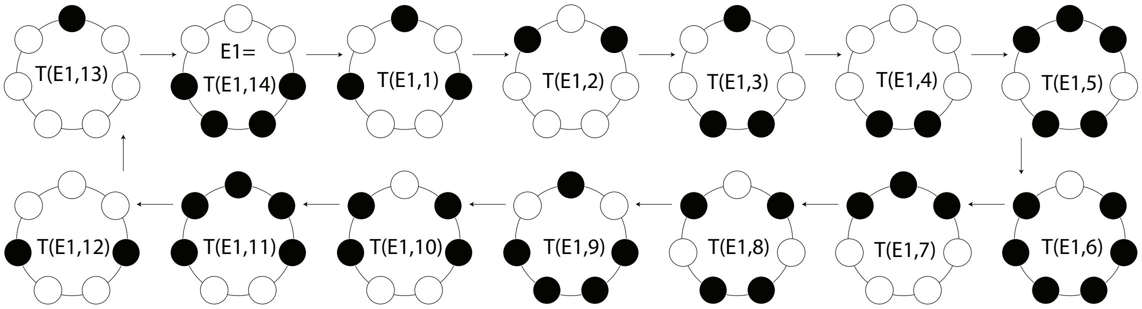
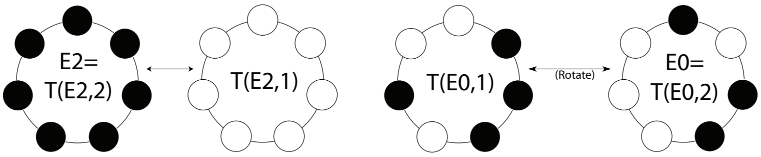


Figure 3. Successive states of E_0 and E_2 under T.



We now compute the values $T(E_0, n)$ and $T(E_1, n)$ by applying the evolution rule. The results are seen in Fig. 2 and Fig. 3. Note that $T(E_0, n) = T(E_0, n+2)$ (keeping in mind that we consider configurations that differ only by a rotation or reflection to be identical) and $T(E_1, n) = T(E_1, n +$

$14)$. This cyclic nature of the context under time development will allow us to find the (overall) self-replicability, rather than just the self-replicability over some constrained time.

Next, we calculate the 'presence' of the subsystem S (see Fig. 1b) in each of $T(E_{0,1}, t)$ with $t \in \{0, 1, 2, 3, ..., 13\}$, by counting how many times it occurs in each time step. The results are presented in Table. 1.

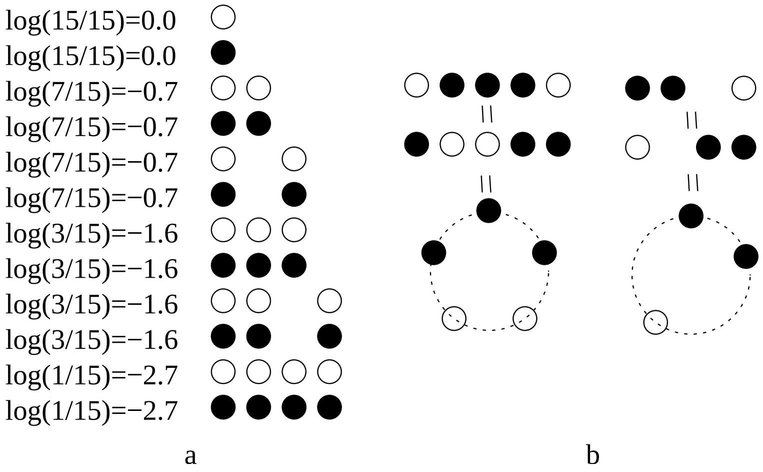
Table 1. Cellular Automaton system: Time and Presence of S in E_0, in E_1, and in E_2.

(t)	$P(T(E_0, t), S)$	$P(T(E_1, t), S)$	$P(T(E_2, t), S)$
0	7	0	1
1	0	2	0
2	7	0	1
3	0	0	0
4	7	0	1
5	0	0	0
6	7	1	1
7	0	4	0
8	7	1	1
9	0	0	0
10	7	2	1
11	0	0	0
12	7	3	1
13	0	0	0
Totals:	49	13	7

We now take the log of the quotient of the sums, yielding $\log(13/7) = 0.619$ for the self-replicability over $t = 0...13$. Note that since T is cyclic, the self-replicability from $t = 0$ to $t = 14n - 1$ is, for any integer n, $\log(\frac{n*\sum_1}{n*\sum_2}) = \log(\frac{\sum_1}{\sum_2})$ where $\sum_i$ denotes $\sum_{t=0}^{13} P_O(T(E_i, t), S)$. It is easy to see that the overall self-replicability will converge to this value as well. All the self-replicabilities for the possible systems in this context are calculated in an identical fashion. In Fig. 4, we show the replicabilities of all systems in a similar, 5-node ring in a context containing the given system relative to a 'blank' context.

It is interesting to note that by 'simply' changing the size of the context, we no longer have a positive value for the 'three adjacent nodes filled' case. Indeed, as Fig. 4 lists all subsystems for which replicability is defined, we have lost the ability to even have positive values. Even small changes in a context can significantly alter the patterns that can be observed as that context develops, which supports the point that replicability should not be looked at merely as the property of a system, but as a function of both a system and its context.

Figure 4. Cellular Automaton system: (a) *Overall* Replicability $R_O(S, E_1, E_2)$ for 'ring of 5' Cellular Automaton example, where $S =$ is given on each line, E_1 is the automaton with exactly two adjacent cells filled, and E_2 is the automaton with all cells hollow. Note that many systems are not ϵ-present in the 2-cycle system, and thus their self-replicability is not defined relative to it. (b) Notation used for depicted subsystems.



4.2. *Ring System*

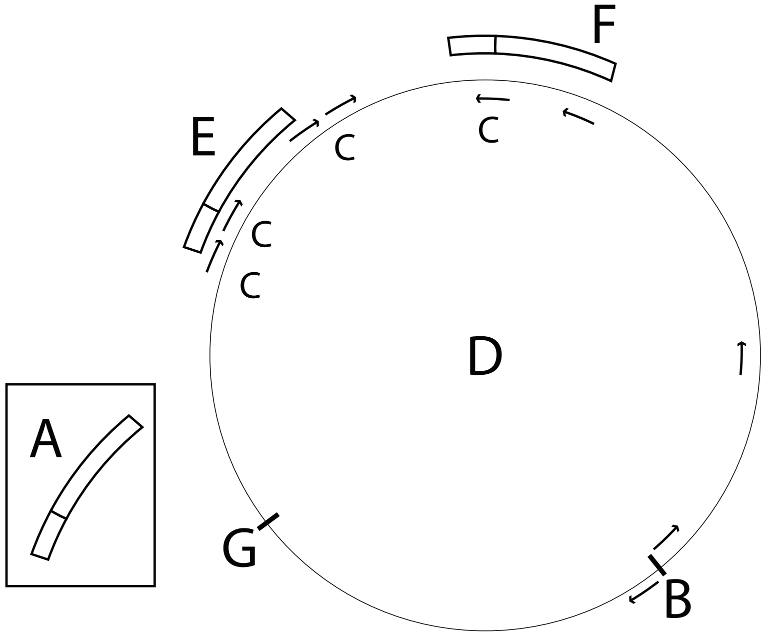
The second system gives an example where the amount of information in the context grows as the time development function is applied. For conceptualization purposes, the system can be seen as an ideal closed fiber-optic ring (Fig. 5D), with a number of irregularities (Fig. 5B, Fig 5G) that act as beam splitters (Fig. 5B), into which a pattern of moving light, packets (Fig 5C) can be injected. We idealize the light packets by ignoring wave dispersion and attenuation, signal loss, and entropy. We assume that light packets travel at a constant speed, clockwise or counterclockwise. For matching patterns, we specify some series of spaces (Fig 5A) and scan the ring for locations where light packets have matching positions (Fig 5E), ignoring any partial matches (Fig 5F).

In the case where each irregularity splits the immediate segment (the part of the ring bounded by the closest irregularities on either side) into segments whose lengths have a rational quotient, any inserted pattern of light packets will reach an equilibrium configuration. This configuration is closely tied to the equilibrium configuration reached by a single injected point, and the self-replicability falls into one of two almost trivial cases:

- First, the system in question may be found in the single-injected-point equilibrium configuration. Here, inserting the system into the ring will be neither better nor worse in the long run than inserting even one light packet, yielding a zero self-replicability.
- Second, the system in question is not found in the single-point equilibrium. Then, the only way for it to arise is to have been specifically placed in the ring, which is a possibility we have excluded from consideration.

For example, we take the following situation. If the ring is assigned a circumference of 2π, so that locations on the ring may be specified by radian measures, and beam-splitting irregularities are placed

Figure 5. Illustration of example Ring System. A: A system with potential self-replicability (a pattern to match) B: An irregularity with recently split packets C: optical packets D: center of the ring E: the system (A) acceptably matching a set of packets, F: the system (A) not acceptably matching a set of packets G: location of a second irregularity.



at locations 0 and π, then wherever we start an intensity 1 light packet, we will achieve an equilibrium state which, at every whole natural time unit (the time it takes for a packet to travel a distance of 2π) will have two light packets of intensity $\frac{1}{2}$, located at the starting point and its antipodal point. (These could also be considered as four light packets of intensity $\frac{1}{4}$, where at the points listed there are two packets traveling in opposite directions which happen to momentarily coincide.)

Since the light packets do not interfere with each other, this result for one packet can be extended to any initial collection of packets, and we note that, in this configuration of beam splitters, the equilibrium state will, at whole natural time intervals, consist of the original collection at $\frac{1}{2}$ intensity, and an antipodal collection at $\frac{1}{2}$ intensity.

If we are trying to find the self-replicability of the pattern “a packet, then a gap of length π, then a packet”, then no matter what collection of packets we begin with, be it the pattern we wish to match, a single packet, or something more involved, the ring will quickly achieve an equilibrium state consisting of collections of the sought-after pattern. This is the first case, where the pattern we search for is the single-injected-point equilibrium configuration.

Note that here we are looking at patterns hosted by the underlying ‘physical’ system. While the light packets are propagating and their population is growing in number, that growth is only of interest insofar as it provides more opportunities for the pattern we are searching for to arise.

In this specific case, if we denote E_1 as the initial ring with the pattern in place, and E_0 as the initial ring with just one light packet, then we will effectively have that the presence if the pattern in E_0 after time $t - 1$ is the same as the presence of the pattern in E_1 after time t. Plugging this relation into the definition for self-replicability, we see

$$R_S(S, E_1, E_0, \epsilon) = \lim_{x \to \infty} \log \frac{\int_{t=0}^{x} P_\epsilon(T(E_1, t), S)dt}{\int_{t=1}^{x} P_\epsilon(T(E_1, t-1), S)dt} \tag{6}$$

This is effectively $\log(x + a) - log(x)$ when x becomes large and where a is a constant, which tends to 0, giving us the self-replicability of the first trivial case.

In the second trivial case, we note that, as mentioned above, the system achieves an equilibrium packet distribution composed of antipodal pairs derived from the original distribution. If we wish to search for a pattern other than the single-injected-point equilibrium configuration, and the original packet distribution does not already contain the pattern we wish to search for, then the pattern we wish to search for will never arise. Since we have not defined replicability in cases where the sought after pattern never arises in one or both contexts, we do not consider this second trivial case.

The problem becomes more interesting when allowing for beam-splitting irregularities that divide the ring into irrational proportions (See Fig. 6 for a variety of divisions). In this case, no steady state exists. While any initial distribution should become densely spread over the ring, the distribution is not necessarily uniform: thus, the presence of a given pattern is possibly nontrivial.

4.3. Exponential Growth System

The third system is a model of a physical system that is modeled abstractly, rather than explicitly simulated. The goal is to measure the self-replicability of an expanding system, satisfying two properties:

- First, once a seed is established, the presence of the system in the context is given as a function of time. For this example, we specify an exponential function of time.
- Second, there is a constant probability rate $c \in (0, 1]$ that the seed will spontaneously appear in a unit period time. The probability that a seed has not spontaneously formed by time t is given by $(1 - c)^t$. This guarantees that a system S is always possible in a context E, and self-replicability thus is defined.

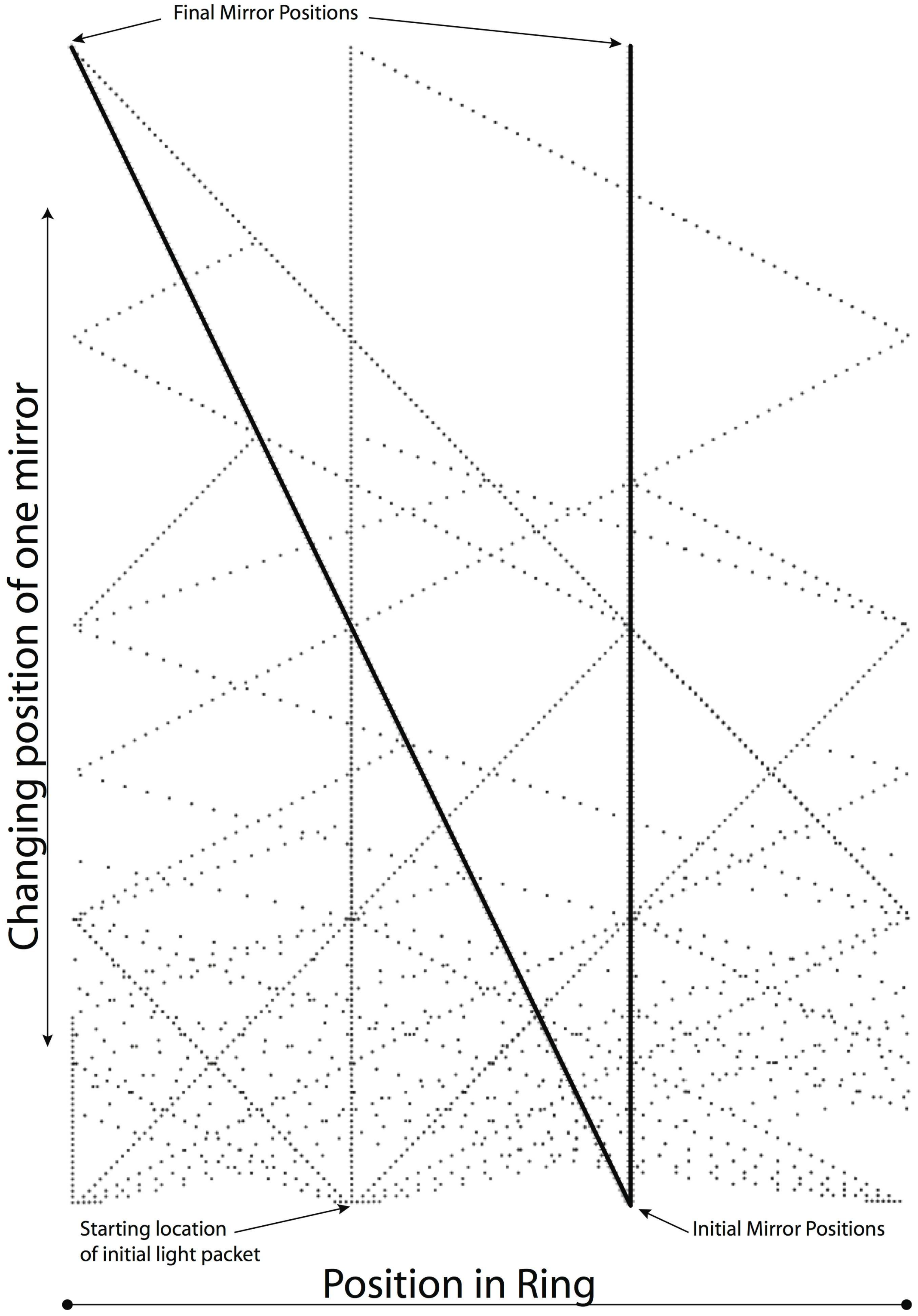
Since we are not going to directly simulate, we do not actually define the dissimilarity pseudo-metric, time development function, system, or contexts. Instead, we define the important relationships that are actually utilized in calculating self-replicability. In particular, we require that the presence (which would be determined by the dissimilarity pseudo-metric and the structure of the context) of system in the context increase exponentially from the time a seed is inserted, i.e.

$$P_\epsilon(T(E_0, t), S) = ke^t \tag{7}$$

Here, k is a normalization constant whose choice depends on the size of the context. Specifically, it depends on the time at which we wish the entire context to become full of instances of the system. Thus, during the period while $ke^t < 1$ we compare a context E_1, seeded with a seed S, to an unseeded, homogeneous context $E_0 = E_1 - S$ and already have a large part of the self-replicability formula:

$$R_S(S, E_0, E_1, \epsilon) = \lim_{x \to \infty} \log \left(\frac{\int_{t=0}^{x} P_\epsilon(T(E_0, t), S)dt}{\int_{t=0}^{x} P_\epsilon(T(E_1, t), S)dt} \right) \tag{8}$$

Figure 6. Ring system: Behavior of the light packet (dot) distribution as mirror locations (lines) change. A family of ring systems have been stacked horizontally with one mirror location fixed (vertical line) and one varying (diagonal line) across the family. Every horizontal cross section illustrates the positions of split-apart light packets after one natural time unit (the time required for a packet to make a full cycle around the ring.) Data derived from computer model (see Appendix 7.).



At this point we pause to emphasize whose presence we are trying to capture. While the hypothetical crystal is growing, we are not seeking to equate physical expansion with replication. Rather, we are interested in the replication of crystalline patterns. The idea is similar to asking "Given an $m \times m$ square grid, how many squares can be drawn along the gridlines?" Increasing the number of gridlines is not of particular interest, but the number of patterns that such an increase allows is. Here, with this crystal model, we are looking at the replication of underlying patterns and the information in them, rather than the physical structure that plays host to those patterns.

We know there is some chance at any time of a seed forming, but we do not know when, exactly, a seed might form. We still treat the underlying system as deterministic, so $T : \overline{E} \rightarrow \overline{E}$ remains a well-defined function. But of all the possible time development functions, we are in the dark as to which one is in use. All we have is a probabilistic condition. To address this, we can use the seed-formation probability to weight the presence. There is a $(1-(1-c)^t)$ probability that a seed *has* formed by time t, so we use this to weight the presence we would have obtained, had a seed formed at a specific time, using the equation

$$P(T(E_1,t),S) = \int_0^t P(T(E_0,t-s),S)(1-(1-c)^s)ds \tag{9}$$

Plots of the values this probability-weighted presence take on can be seen in Fig. 7 in dotted lines, while a plot of the presence given a seed at time 0 is in the same figure with a solid line.

Using this for the presence in $T(E0,t)$, we derive the following integral form for the overall self-replicability, and plot it in Fig. 8:

$$\begin{aligned} R_S(S,E_0,E_1,\epsilon) &= \lim_{x\to\infty} \log\left(\frac{k\int_{t=0}^{x} e^t dt}{\int_{t=0}^{x}\left(\int_{s=0}^{t}\left(P(T(E_0,t-s),S)(1-(1-c)^s\right)ds\right)dt}\right) \\ &= \lim_{x\to\infty}\log\left(\frac{k(e^x-1)}{\int_{t=0}^{x}\int_{s=0}^{t} ke^{t-s}(1-(1-c)^s)dsdt}\right) \\ &= \log\left(1-\frac{1}{\log(1-c)}\right) \end{aligned} \tag{10}$$

Note that although we should confine ourselves to $x < t'$ (where $t' = log(1/k)$ so that $ke^{t'} = 1$), the result will always be independent of the choice of k, so we need not constrain our time interval due to the normalizing factor.

So, we have a result that visibly converges in the long run (two examples seen in Fig. 8), with the magnitude of the self-replicability (in this example, with exponential growth) being inversely proportional to the probability of seed formation. For example, if the constant c is at $c = 0.1$, then $R_S = 2.35$, while when c is much smaller, such as $c = 0.001$, we have $R_S = 6.88$, and a high value (yielding a probability of seed formation), like $c = 0.999$, yields $R_S = 0.13$ (i.e. when it is close to certain a seed will form in any event, knowing we have one for certain is not a big difference, so the self-replicability factor is small.)

Figure 7. Plot of the log of the system presence, $\log(P_{\epsilon}(T(E_1, t), S))$ vs. t for different values of c, the probability of random seed formation. We use the logarithm to make long-term behavior more visible.

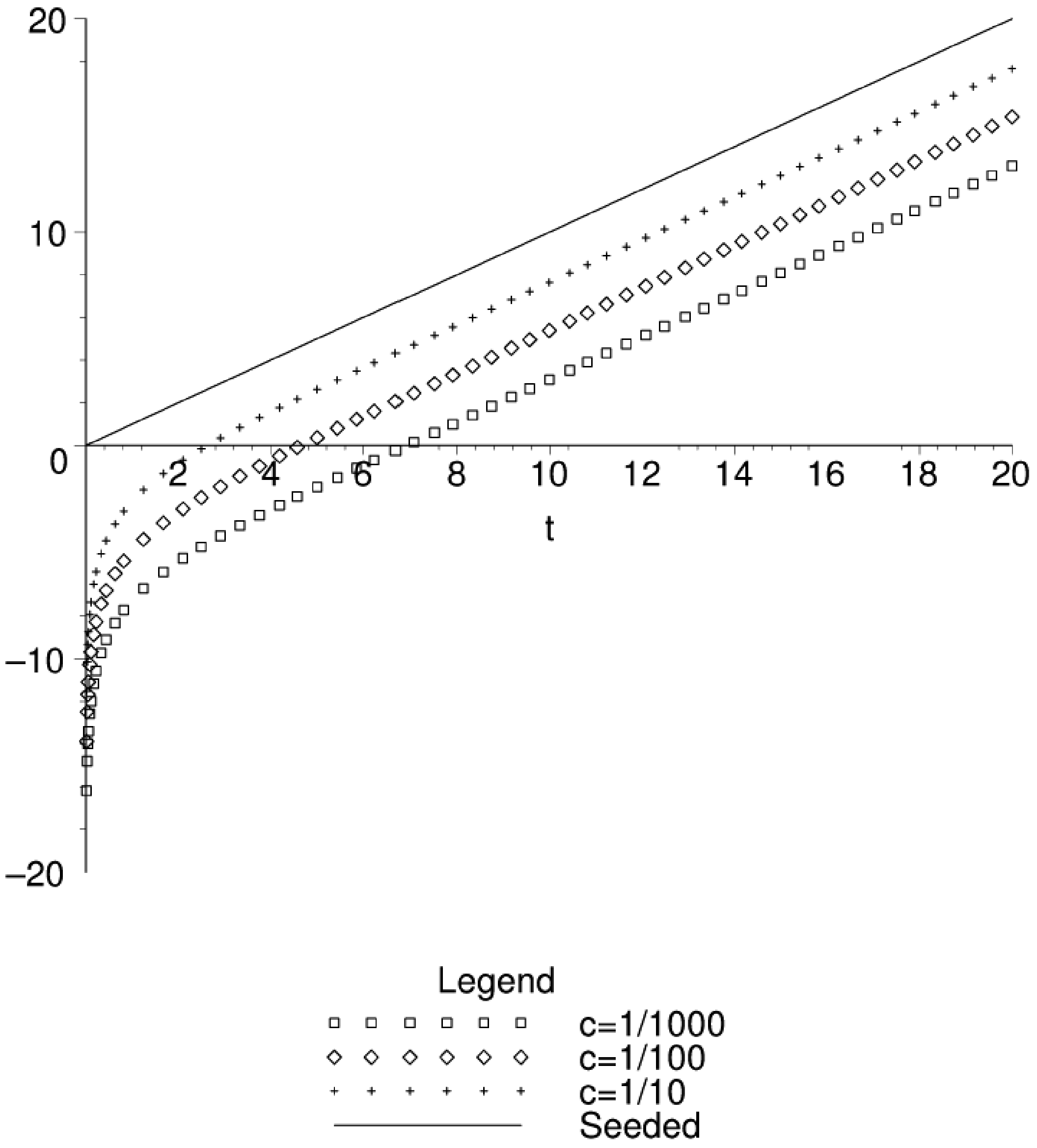
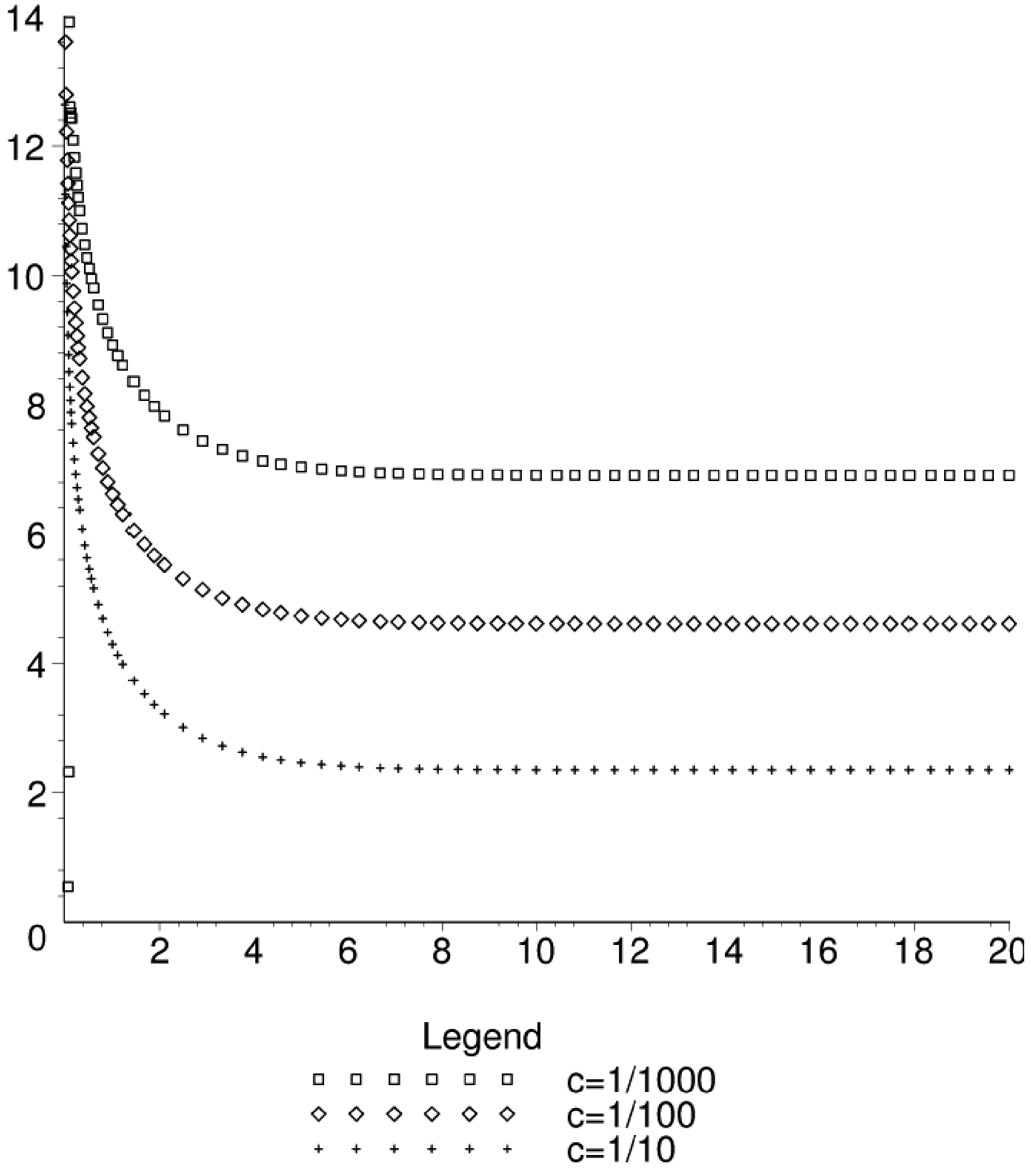


Figure 8. Plot of $R_T(E_1, E_2, S, \epsilon, 0, t)$ vs. t for the system model, showing the same values of c as in Fig. 7. Note that R_T converges quickly (toward R_O) as t becomes large.



4.4. L.S. Penrose's Self-reproducing Machines

For our final system, we consider a particular mechanical self-reproducing machine of L.S. Penrose [19]. These consist principally of a system of releasable latches to hold structure, and a seesaw-like element (henceforth called a tumbler) to carry information. They operate in an essentially linear (assume horizontal) world with some sort of side-to-side agitation.

In Penrose's original conception, "the units... were to reproduce themselves only when the object to be replicated was introduced as a pattern for copying." He wished to ensure a "no life except from life" condition [19]. With only this statement, it is clear that as presented, Penrose's machines would have infinite self-replicability: If one is not present in a context, none ever will be, and the denominator in the relevant equation will be zero.

For a nontrivial variant, our model (see appendix 7.) incorporates a nonzero probability of two 'neutral' units colliding to form a replication-capable pair. Since Penrose's machines ensured this would not happen by properly squaring off the ends of the tumblers, our conceptual variation can be seen as introducing a variable degree of roundness to the ends of the tumblers. The closer to square the ends of the tumblers, the lower the probability of a 'wobble' that would allow two neutral units to join. The less square the ends, the higher the chance of a join-permitting wobble. For our model, we assumed the rounding was symmetric, so that the chance of a wobble inducing a left-tilted pair was equal to the chance of producing a right-tilted pair.

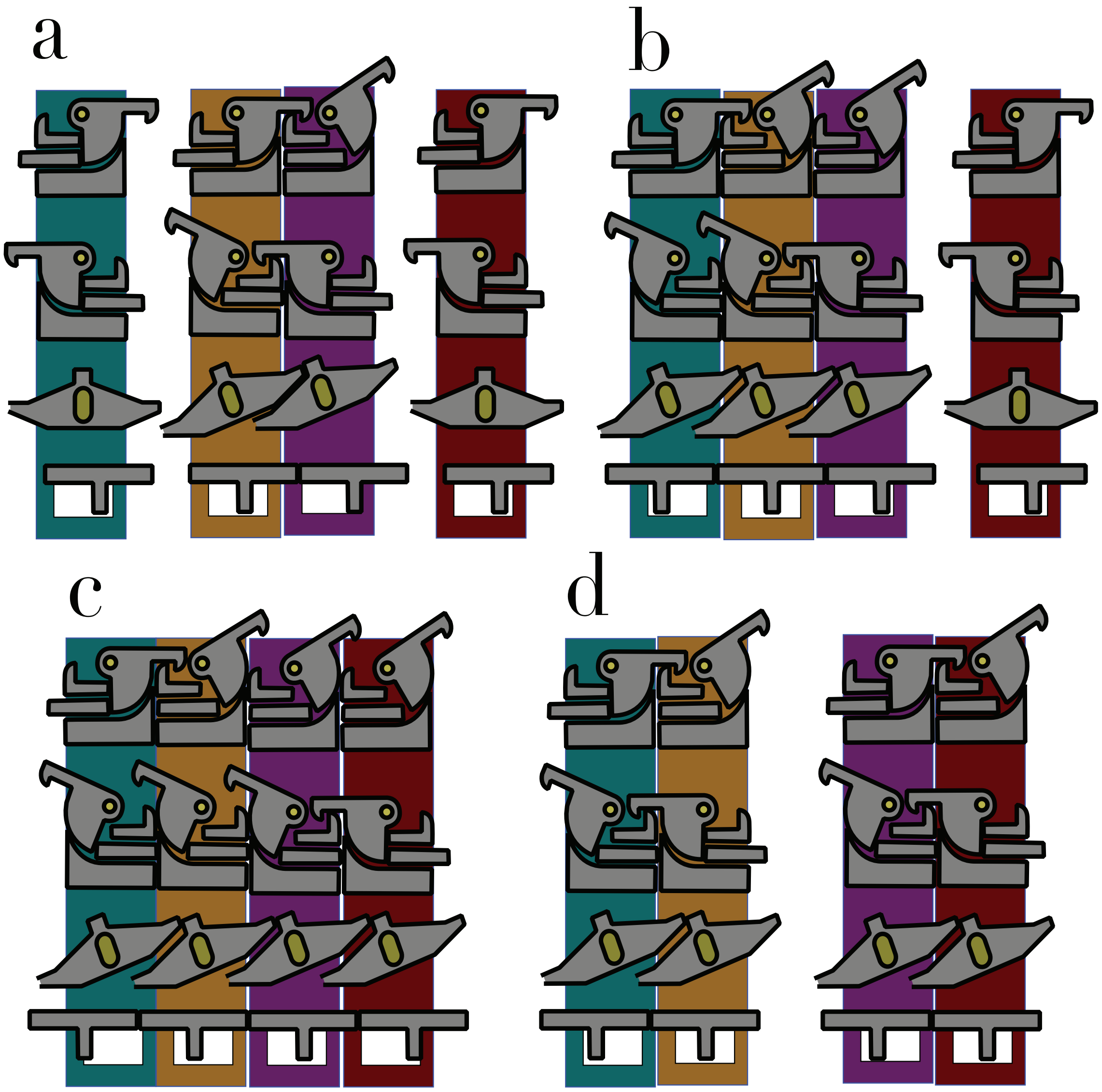
We modeled these machines in software, and for horizontal perturbation assumed that all pieces except one began at rest, and one piece was introduced from the left with some velocity. A series of collisions would propagate through the units in turn and, upon reaching the last unit, would be reflected back in the opposite direction. Atomic units (a tower consisting of a pair of releasable latches and a tumbler) could be in one of three states: tilt-left, tilt-right, and tilt-neutral. For our tumbler-rounding factor, we assumed a 1% chance per collision of a wobble occurring, which then had an even probability of becoming tilt-left or tilt-right.

Given various population sizes and initial tilt-configurations of atomic units, two measurements were made based on averaging over at least 1000 simulations per initial condition: the time required to reach an equilibrium with only tilt-right and tilt-left groups (which cannot change without the presence of a tilt-neutral resource), and the relative population of tilt-right and tilt-left groups. The data are available in the appendix.

As no changes could occur after reaching an equilibrium state, it is possible to exactly derive self-replicability (see appendix for details) which in the case of Penrose's machines is between $\log(1)$ and $\log(2)$, depending on which initial configuration we choose as our context. These extremes are realized for large populations by, first, the case where the seed population is at the far right end of the row (recall that collisions begin at the left side and propagate rightward), which gives rise to a $0 = \log(1)$ replicability (in the long run, no better than counting on a wobble to start reproduction), and second, the case where the seed population is at the far left end of the row, which guarantees that all neutral units will be used to reproduce the seed population.

This model adds in a random chance of spontaneous wobbles and uses a serial collision model instead of a random-agitation one, as in Penrose' formulation. It only simulates a single-rocker Penrose machine, though the model could be expanded to arbitrary rockers with negligible execution overhead. It does restrict attention to the linear case that Penrose begins with, rather than attacking the two-dimensional extension which, going by Penrose's diagrams, we believe would have lower replicability: they use "inter-digitating baseplates" which appear as if they would require extremely unlikely approach vectors to actually couple, a problem not faced by the 1-D system. However, by engineering the base plates to make an acceptable meeting more likely, the replicability should be able to be improved on, with the 2-D case as a (probably not least) upper bound.

Figure 9. Penrose's self-reproducing machines (a) with one reproducing pair of tilt-left units present in the center and two tilt-neutral units on its sides, (b) during an initial reproduction phase, (c) at the end of reproduction, and (d) after splitting to complete the reproduction. These would have been represented as $[U, LL, U]$, $[LLL, U]$, $[LLLL]$, and $[LL, LL]$ respectively within the computer model.



5. Recap

The cellular automaton gave rise to systems with negative self-replicabilities, relative to the 'empty' state, and systems with undefined self-replicability, which would never arise from the empty state. In essence, any pattern that was viable in the two types of cycles turned out to be more viable in the simple, all-filled/all-hollow cycle. Doing the calculations to examine the self-replicability of a pattern in E_1 relative to E_2 where both contexts can be found in the more complicated six-cycle yields zero self-replicabilities. This is as expected, since the only difference between different steps in the six-cycle is time, and the action of the time development function is transitive.

The light-ring example gave a view of one possible physical system with nontrivial, positive self-replicability. It also demonstrates the limits of simulation-based calculations. It is possible to achieve some results by this method; however, the quantity of data involved in direct measurement becomes quickly unmanageable for even a very simple continuous system. Additionally, the light-ring example demonstrates low self-replicability arising from a system whose replication is extremely dependent upon the workings of the context. The values found when the flaws split the ring into irrational components were nonzero but low, and when the components were in rational proportions, the self-replicability (when defined) was simply zero.

The use of abstract methods to calculate self-replicability in the growth model, while constrained in accuracy by the quality of the model being used, essentially reduces the problem to the level of symbolic and/or numeric integration. Since the dissimilarity pseudo-metric is required only for constructing a presence function, by moving directly to the presence function we can skip defining the pseudo-metric. Similarly, the heavy computations involved in applying the time development function are subsumed in the constraints placed on the time-developed behavior of the presence function. Although these are still inherent parts of the concept of replicability, it is important to note that they need not be explicitly given in the calculation of self-replicability.

In some cases, however, explicit calculations are possible: Given an appropriate model which, in particular, makes information about the possibility of 'spontaneous generation' accessible, approximations can at least be calculated for self-replicability. In the case of measuring Penrose's self-reproducing machines, the design was well specified enough to provide clear asymptotic behavior and, from that, average measurements for replicability.

The flexibility of calculation reflects the idea that replication is a property of the information tied up in the interaction between a system and its context, rather than in either one or in the specific physical makeup. Also, the growth model in particular demonstrates how self-replicability succeeds in coinciding with intuition: The more likely it is for a system to manifest regardless of having been present before, the lower its self-replicability will be; conversely the more unusual and unlikely a system is which is capable of replicating its existence, the higher its self-replicability.

6. Replicability Comparisons

Faced with a complex self-replicating system with many internal degrees of freedom, operating in a context as rich as the laws of physics, the task of even finding a model to use for calculating an approximate number for self-replicability is daunting, and data on the probability of biological life arising in a context absent of pre-existing life is hard to come by. The importance of a measure is not so much the particular numbers it gives to any particular system, but the ability to make comparisons between systems (or compare different contexts for a given system).

As such, even if deriving specific numbers for a system is intractable, this kind of multiply-valued measure of replicability could be directly useful if comparisons could still be made. Take, for example, the 'Molecubes' of Mytilinaios, Marcus, Desnoyer, and Lipson [20]. These are cubical units with controllable electromagnets on each face and a swivel on a diagonal (splitting the cube roughly into two tetrahedral halves). A linear arrangement of just four Molecubes may, by turning on and off one of its electromagnets and going through a series of deformations, construct an identical linear arrangement by drawing cubes from a fixed-location feed.

Consider this compared to a linear arrangement of, say, eight Molecubes, which could perform a similar self-reproduction. Set in identical contexts, operating under all the same physical laws, the only difference between these two situations is the number of basic components constituting them (and the control program, which we ignore for the time being.) Giving the larger Molecube replicator the benefit of the doubt by assuming both systems replicate at the same rate (this could be achieved by throwing time-wasting instructions into the smaller replicator's control program), the only difference that would arise when calculating their self-replicabilities would be that the 8-cube system is less likely to arise from randomly throwing basic building blocks together than the 4-cube system is.

That difference indicates that the bottom term in the self-replicability formula will be greater for the 4-cube system than the 8-cube system, so the whole quotient will be smaller for the 4-cube than the 8-cube, so the self-replicability of the 4-cube system will be strictly less than that of the 8-cube system. That is to say, all other things being equal, the more complex (that is, the less likely to by spontaneously assembled) system will have higher self-replicability.

Of course, this assumed that the replication speed was the same for both systems, which is very unlikely. Again, however, at least one comparison can be made, this time between a 'slow' 4-cube system and a 'fast' 4-cube system. In a similar manner, since the top term in the self-replication formula would be at least as large for the fast system as the slow system, the replicability of the fast system would be at least that of the slow one. All other things being equal, a faster replicator will have higher self-replicability.

While it is seldom the case that all other things actually will be equal, our conception of self-replicability allows at least some such comparisons without *any* calculations, and provides a framework in which more complicated comparisons can be made which need not require complete calculations of replicability.

7. Conclusions and Future Work

We have proposed a means of measuring the fundamental property of self-replication, seeking a graded measure that can be universally applied to systems from cellular automata to living systems. We provided four examples that involved various mixtures of discrete and continuous variables and were handled both through direct simulation and analytical modeling. This property is not simply seen as intrinsic to a system, but is found in the information that arises from the interaction between a system and its context.

We currently see two branches of further immediate investigation:

On the theoretical side, while a few properties, such as additivity in Equation 4, are known, other properties, such as continuity of the replicability function with respect to tolerance, warrant further investigation. Given continuity in the choice of tolerance, exact measures of replicability could be derived as a limit of approximations using successively smaller error tolerances. Additionally, identifying *all* factors with which, "if all else is equal", self-replicability varies monotonically should allow for comparative arguments which do not require full calculation of the exact self-replicability, but rather only require knowing information like "this machine uses fewer distinct fundamental building blocks" or "the fundamental components of this reproducer have lower complexity". Past that, a natural step is trying to get a handle on situations where "all but N things are equal" which allow property-wise comparisons without extensive calculations .

In addition to further investigation of self-replicability, we wish to develop and use the additional frameworks introduced in the definitions section to use the more general notion of replicability (*sans* 'self') to make comparisons between systems and comparisons between contexts. The definitions already provide relative replicability for comparing the performance of a given replicator in different contexts, as long as those contexts both are states of the same environment-system. In order to provide 'comparative replicability' between different systems in the same environment, more development on the $E - S$ 'unseeding' notion should allow for a well-defined notion that parallels the definition of relative replicability.

On the applied side, it would be desirable to calculate replicability in more intricate systems. Such systems range from continuous self-replicating automata [21], simple auto-catalyzing enzymes (prions), and molecules [22] to robots [15], and even factories [23]. Currently, significant biological systems seem intractably complex, but establishing monotonic factors and tolerance-continuity could allow comparison to, and between, biological systems. Given our view that self-replication is an essentially information-centric property, calculations for social 'memes' such as 'universal suffrage' or ' adoption of new technologies (such as cellphones)' would be particularly interesting.

Again, moving beyond self-replicability, the theoretical work on 'comparative replicability' would allow for practical application of these measures. For example, Lee, Moses, and Chirikjian recently announced additional self-replicating Lego robots[17] and ranked them using measures based on entropy and based on the quantity of active elements and their interconnections. These measures, like self-replicability, assign a particular value to each system, and the values are subsequently compared. With a direct 'comparative replicability' measure, however, systems could be ranked relatively without having to establish their absolute value. In cases where systems have very small self-replicabilities, direct comparison should be less susceptible to numerical rounding error than calculating and then comparing self-replicability.

Ultimately, we seek to get a better idea of the scale into which interesting replicabilities fall, methods of executing low-computation comparisons, and the conditions under which self-replication is maximized (both for systems and for contexts).

Acknowledgements

This work was supported by the U.S. Department of Energy, grant DE-FG02-01ER45902.

References

1. Sipper, M.; Reggia, J. Go forth and replicate. *Scientific American* **2001**, *285*, 34–43.
2. Nehaniv, C.; Dautenhahn, K. Self-Replication and Reproduction: Considerations and Obstacles for Rigorous Definitions. In *Third German Workshop on Artifial Life: Abstracting and Synthesizing the Principles of Living Systems*; Wilke, C.; Altmeyer, S.; Martinetz, T., Eds. Verlag Harri Deutsch, Thun and Frankfurt am Main, 1998, pp. 283–290.
3. Adams, B.; Lipson, H. A Universal Framework for Self-Replication. In *Advances in Artificial Life: 7th European Conference, ECAL 2003 Dortmund, Germany, September 14-17, 2003 Proceedings*. Springer, 2003, p. 1.
4. Moore, E.F. Machine Models of Self-Reproduction. In *Essays on Cellular Automata*; Burks, A.W., Ed. University of Illinois Press, 1970, pp. 187–203.
5. Lohn, J.D.L.; Reggia, J.A. Automatic Discovery of Self-Replicating Structures in Cellular Automata. *IEEE Trans. on Evolutionary Computation* **1997**, *1*, 165–178.
6. Von Neumann, J. Von Neumann's Self-Reproducing Automata. In *Essays on Cellular Automata*; Burks, A.W., Ed. University of Illinois Press, 1970, pp. 4–65.
7. Langton, C.G. Self-Reproduction in Cellular Automata. *Physica D* **1984**, *10*, 135–144. *Details of a simple self-reproducing CA configuration.*
8. Păun, G. Computing with membranes. *Journal of Computer and System Sciences* **2000**, *1*, 108–143.
9. Löfgren, L. An axiomatic explanation of complete self-reproduction. *Bulletin of Mathematical Biology* **1968**, *30*, 415–425.
10. Rosen, R. On a logical paradox implicit in the notion of a self-reproducing automation. *Bulletin of Mathematical Biology* **1959**, *21*, 387–394.
11. Guttman, B. A resolution of Rosen's paradox for self-reproducing automata. *Bulletin of Mathematical Biology* **1966**, *28*, 191–194.
12. McMullin, B. John von Neumann and the Evolutionary Growth of Complexity: Looking Backwards, Looking Forwards. *Artificial Life* **2000**, *6*, 347–361.
13. Sanchez, E.; Mange, D.; Sipper, M.; Tomassini, M.; Pérez-Uribe, A.; Stauffer, A. Phylogeny, Ontogeny, and Epigenesis: Three Sources of Biological Inspiration for Softening Hardware. In *Evolvable Systems: From Biology to Hardware, First International Conference, ICES 96, Tsukuba, Japan, October 7-8, 1996, Proceedings*; Higuchi, T.; Iwata, M.; Liu, W., Eds. Springer, 1996, Vol. 1259, *Lecture Notes in Computer Science*, pp. 35–54.
14. Pati, A.K.; Braunstein, S.L., Can Arbitrary Quantum Systems Undergo Self-Replication. World Scientific, 2008, pp. 223–229.
15. Chirikjian, G.; Zhou, Y.; Suthakorn, J. Self-replicating Robots for Lunar Development. *IEEE/ASME Transactions on Mechatronics* **2002**, *7*, 462–472.
16. Hornby, G. Toward the Computer-Automated Design of Sophisticated Systems by Enabling Structural Organization. In *Symposium on Complex System Engineering '07*, 2007.
17. Lee, K.; Moses, M.; Chirikjian, G.S. Robotic Self-replication in Structured Environments: Physical Demonstrations and Complexity Measures. *The International Journal of Robotics Research* **2008**, *27*, 387–401.

18. Zykov, V.; Mytilinaios, E.; Adams, B.; Lipson, H. Self-reproducing machines. *Nature(London)* **2005**, *435*, 163–164.
19. Penrose, L.S. Self-Reproducing Machines. *Scientific American* **1959**, *200*, 105–114.
20. Mytilinaios, E.; Marcus, D.; Desnoyer, M.; Lipson, H. Designed and Evolved Blueprints for Physical Self-Replicating Machines. In preparation, 2004.
21. Smith, A.; Turney, P.; Ewaschuk, R. JohnnyVon: Self-Replicating Automata in Continuous Two-Dimensional Space. Technical Report ERB-1099 (NRC 44953), National Research Council, Institute for Information Technology, 2002.
22. Freitas, Jr, R.A.; Merkle, R.C. *Kinematic Self-Replicating Machines*. Landes Bioscience: Georgetown, TX, 2004.
23. Freitas, R.; Zachary, W.; Gilbreath, W. A self-replicating, growing lunar factory. In *Space Manufacturing - Proceedings of the Fifth Princeton/AIAA/SSI Conference on Space Manufacturing*; Grey, J.; Hamdan, L.A., Eds., 1981, pp. 109–119.

Appendix A: Modeling Penrose's machines

A symbolic model of the atomic units pictured in Fig 9 was executed in SML/NJ. Using syntactical rules applied in a particular order to simulate the effects and order of collisions, respectively. Atomic units were represented by symbols L, U, and R (left-tilt, un-tilted, and right-tilt), and the entire collection was a list of valid strings $\{U, LL, LLL, RR, RRR\}$ identifying the state of all atomic units and all connected blocks. Transition rules such as $(U, LL) \rightarrow (LLL)$ and $(LL, RR) \rightarrow (LL, RR)$ were applied from left to right until reaching the end of the list, then the resulting modified list was reversed and the transition rules were applied again. This was repeated until a fixed state was reached. At that point, the number of collisions that had occurred and the R population relative to the L population would be output.

To add the possibility of a 'wobble', within the transition rules, the (U, U) rule's result was selected randomly, resulting in (U, U) in 99 out of every 100 cases, (LL) in 1 out of every 200 cases, and (RR) in 1 out of every 200 cases.

The initial cases tested were $[R, U, U, , U, U]$, $[U, U, .., R, ., U, U]$ (equal length strings of U on either side), $[U, U, ., U, U]$, and $[U, U, .., U, R]$, referred to as left-start, middle-start, empty-start, and right-start, respectively. Each of these cases were tested for population sizes of 25, 50, 100, 200, and 400, averaging results over a minimum of 500 trials. The data collected are in tables 2 and 3.

Table 2. Penrose simulation data: Number of collisions

Pop size	Left-start	Middle-start	Right-start	Empty-start
25	24	42	46	150
50	49	85	96	166
100	99	172	196	322
200	199	347	395	560
400	399	696	793	1040
Asymptotic N	N-1	$3N/2$	$2N$	$< 3N$

The asymptotic values are based on analysis of the collisions required under the assumption that N is large enough for a 'wobble' to occur somewhere on the first pass and that, unless forced by a known-tilt element, wobbles will give rise to (on average) an even population distribution.

To calculate self-replicability, the various non-empty starts were compared to the empty-start context. Since, after a finite time $t \geq T$, the population size $pop(t)$ became fixed at some value $P = pop(T)$, the integrand in the self-replicability equations became $(\int_{t=0}^{t_1} pop(t) + \int_{t=t_1}^{t_2} P) = (\int_{t=0}^{t_1} pop(t) + P(t_2 - t_1))$. Considering the limit of log of the quotient of two such terms, with P_1 (the fixed population size for one non-empty start) in the numerator and P_2 (the fixed population size for a second non-empty start) in the denominator, the finite parts become insignificant, and we are left with simply $\log(P_1/P_2)$, thus allowing replicabilities to be calculated directly from data, given the context in which to start.

Table 3. Penrose simulation data: R-L Distribution

Pop size	Left-start	Middle-start	Right-start	Empty-start
25	100%	98%	95%	50%
50	100%	97%	89%	50%
100	100%	92%	80%	50%
200	100%	90%	72%	50%
400	100%	85%	62%	50%
Asymptotic N	100%	75%	50%	50%

Appendix B: Modeling the Ring System

For this example, a direct computer model was used, storing the initial conditions (initial light configuration and irregularity locations) and applying the time development rules to determine the location, direction, and intensity of all resulting light packets. Under the assumption that light propagates at a constant velocity, the position and number of split-apart light packets after a unit of time passed could be determined algebraically as a function of packet location, packet direction, and mirror locations.

While running the model by iterating these computations was relatively efficient, the number of packets that had to be tracked still rose exponentially, quickly exhausting computational resources. While some memory was conserved by adding the assumption that packets traveling in the same direction with sufficiently close proximity merged back into a single packet, along with a relaxation of pattern-matching exactness, our re-sources were still not sufficient to run enough time steps to suggest a long term trend. However, whether a trend was indicated or not, All R_S values were no greater than 1 in absolute value. Moreover, those values that appeared to be converging had a self-replicability factor between 0.5 and 0.9.

List of Figures