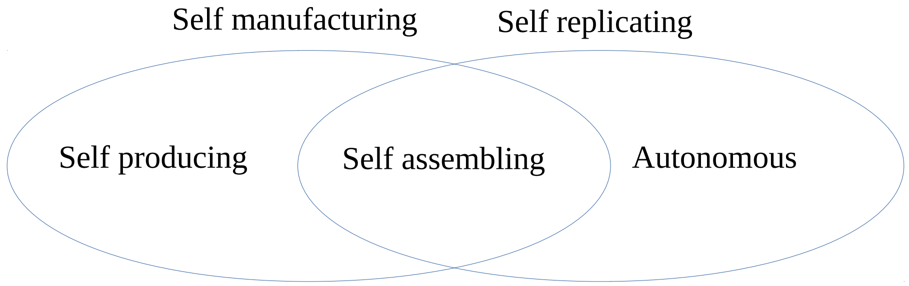
Self Producing & Self Assembling

by Sven Nilsen, 2017

A self manufacturing system is an overlapping concept of a self replicating system. What is common for these two kinds of systems is the ability to self assemble, which is the capability of composing material in physical configurations that produces a copy or functionally equivalent system in capabilities. This give rise to the characteristic property of exponential growth.



In the diagram above, the smallest type of self manufacturing and self replicating systems are depicted. Notice that a self replicating system can be self manufacturing and vice versa. In some contexts, these types can be considered larger. For the scope of this paper, the minimum types are used.

Technically, there exists no pure self producing system, because all systems require energy. A system is regarded as self producing when the material input consist of abundant resources, or the access to material resources is unrestricted. The definition of a self producing system spans the context where the condition of access to material is satisfied. One can often split the lifetime of a system into phases which depends on usage of materials, where each phase is analyzed separately.

The autonomous property of a system is in this context what drives decision making about the process of self replication. This capability can span from very simple control mechanism to sophisticated artificial intelligence.

Self producing and self assembling machine systems are two simple properties of self manufacturing systems. A self producing system contains machines that are made of parts which the machines can produce. A self assembling system contains one or more machines that as a whole can assemble all machines required to produce a copy of the whole system.

```
produces_self(ms: [Machine]) = ∀ i, j { ∃ k, l { ms[k].produces[l] == ms[i].parts[j] } }
assembles_self(ms: [Machine]) = ∀ i len(ms) { ∃ j, k { ms[j].assembles[k] == i } }

Machine : {parts : [MachinePart], produces : [MachinePart], assembles : [nat]}
```

These properties are mathematically defined assuming path semantics, where the “real” machine is externally defined function using the theoretically defined functions are predictors. The predictors are asymmetric paths:

```
real_produces_self[sensory_input → id] <=> produces_self
real_assembles_self[sensory_input → id] <=> assembles_self

sensory_input : [RealMachine] → [Machine]
```

This is a path between functions, so there are examples of systems which satisfies the criteria of self producing and self assembling but do not intuitively satisfy the intention of these definitions. Any path between functions allows too much wiggle room for a precise definition, but it narrows down the behavior such that it can be reasoned about without knowing the whole underlying system embodied in the real world.

In simulations, one can substitute `real_*` and `Real*` with `simulated_*` and `Simulated*`.

The distinction between machines and machine parts is useful because no machine part can assemble anything or produce a machine part. By default, any machine part is considered safe. It is the composition of machine parts into machines and their interactions that give rise to unsafe behavior. Unsafe substances should be reasoned about separately, therefore safety analysis of self manufacturing system assumes safe handling of substances.

A self manufacturing system is considered safe regarding human control, because it requires adding autonomy by e.g. composing with human or algorithmic decision making to form a self replicating system.

A self replicating system is considered unsafe regarding human control, either because of potential conflicts of interests between the human controller and other humans, or competition for resources between a fully autonomous system with low human utility preferences and life forms in general. When designing a self replicating system, is not sufficient to prove non-conflict between an autonomous system and direct human goals in the short term. If any autonomous system does large scale damage to any kind of life form, one should assign a very low utility score to that potential future in general, because it increases the chance that more severe effects are observed.

The unsafeness level of a self replicating system depends on the system’s capacity to acquire parts required for self assembling. Better capabilities means higher unsafeness, unless proven safe.