Consciousness and Self-improving Agents

Department of Philosophy
Central South University
xieshenlixi@163.com
github

October 19, 2019

Figure: One can imagine a detailed floor plan of a room, sitting on a table in the room; this plan has an image of the table on which there is an image of the plan itself. Now introduce the dynamical aspect: the items on the plan are cut out from paper and can be moved to try a different furniture arrangement; in this way the plan models possible states of the world about which it carries information.

The brain contains inside a map of itself, and some neural information channels in the central neural system:

- carry information about the mind itself, i.e., are reflexive;
- are capable of modelling states of the mind different from the current one, i.e., possess a modelling function;
- can influence the state of the whole mind and through that, the behavior, i.e., possess controlling function.

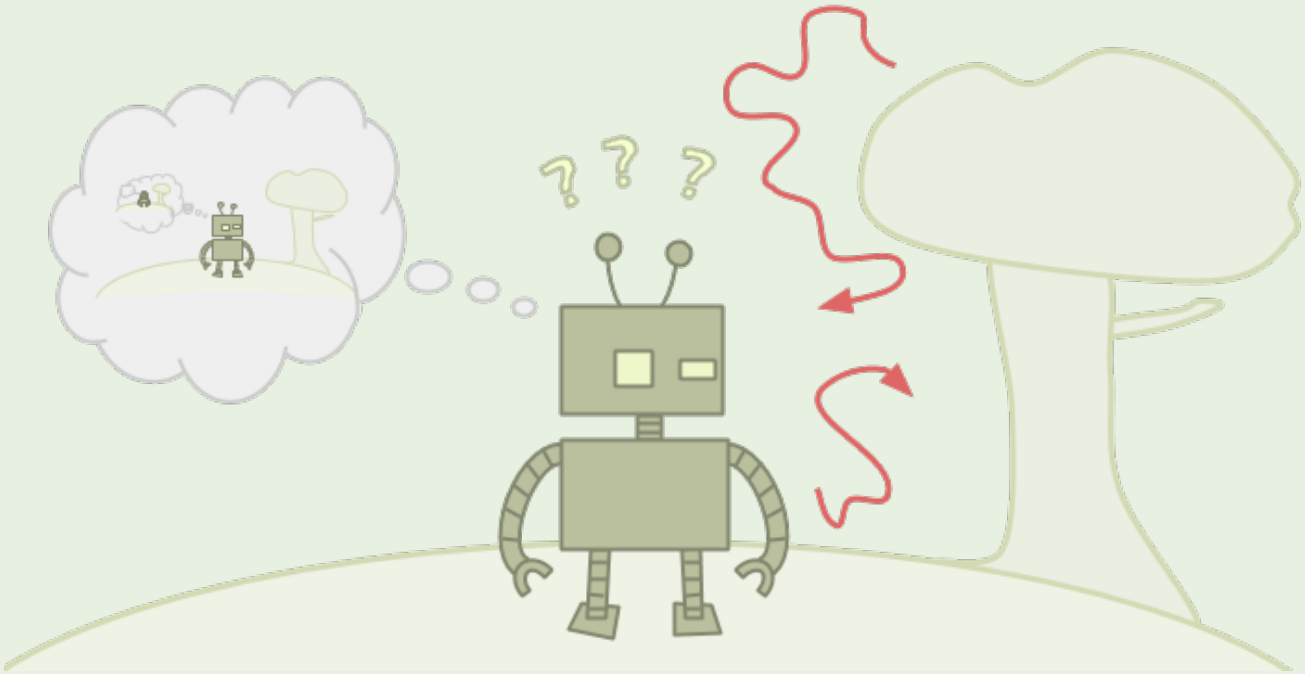
The reflection of the brain inside itself must be coarse grained.

侯士达——“我”是个怪圈

- 有没有意识取决于在哪个层级上对结构进行观察。在整合度最高的层级上看，大脑是有意识的。下降到微观粒子层面，意识就不见了。
- 意识体是那些在某个描述层级上表现出某种特定类型的循环回路的结构。当一个系统能把外部世界过滤成不同的范畴、并不断向越来越抽象的层级创造新的范畴时，这种循环回路就会逐渐形成。
- 当系统能进行自我表征——对自己讲故事——的时候，这种循环回路就逐渐变成了实体的“我”——一个统一的因果主体。

Liar Paradox vs Quine Paradox

- **这句话有 2 个‘这’字，2 个‘句’字，2 个‘话’字，2 个‘有’字，7 个‘2’字，11 个‘个’字，11 个‘字’字，2 个‘7’字，3 个‘11’字，2 个‘3’字。**
- **我在说谎。**
- **把**“把中的第一个字放到左引号前面，其余的字放到右引号后面，并保持引号及其中的字不变得到的句子是假的。”**中的第一个字放到左引号前面，其余的字放到右引号后面，并保持引号及其中的字不变得到的句子是假的。**



Diagonalization

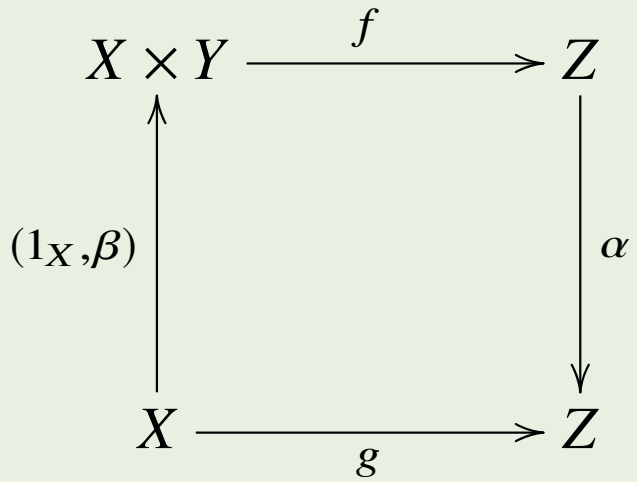
- A function $g\colon X \to Z$ is *representable* by $f\colon X \times Y \to Z$ iff

$$\exists y \in Y \forall x \in X\colon g(x) = f(x, y)$$

Theorem (Lawvere's Fixpoint Theorem)

For sets X, Y, Z, functions $\beta\colon X \to Y$, $f\colon X \times Y \to Z$, $\alpha\colon Z \to Z$, let $g := \alpha \circ f \circ (1_X, \beta)$. Assume β is surjective.

1. *If α has no fixpoint, then g is not representable by f.*
2. *If g is representable by f, then α has a fixpoint.*

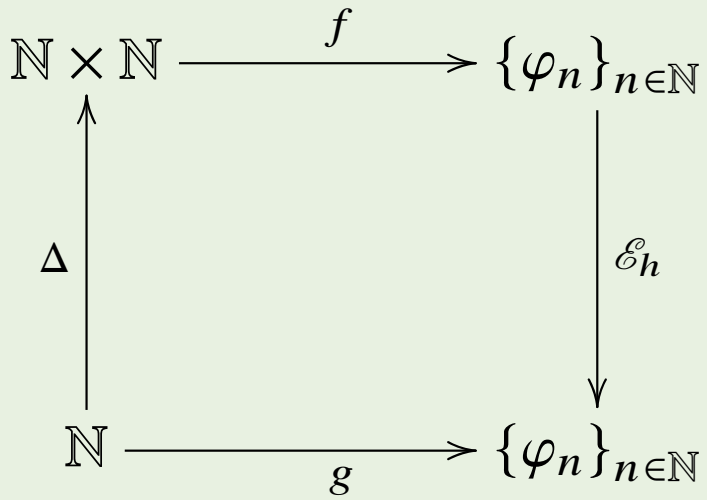


Kleene's Fixpoint Theorem

Theorem (Kleene's Fixpoint Theorem)

Given a recursive function h, there is an index e s.t.

$$\varphi_e = \varphi_{h(e)}$$



where $f : (m, n) \mapsto \varphi_{\varphi_n(m)}$, and $\mathscr{E}_h : \varphi_n \mapsto \varphi_{h(n)}$.
The function $g : m \mapsto \varphi_{h(\varphi_m(m))}$ is a recursive sequence of partial recursive functions, and thus is representable by f. Explicitly,

$$g(m) = \varphi_{h(\varphi_m(m))} = \varphi_{s(m)} = \varphi_{\varphi_t(m)} = f(m, t)$$
$$e := \varphi_t(t)$$

von Neumann's Self-reproducing Automata

Corollary (von Neumann's Self-reproducing Automata)

There is a recursive function φ_e s.t. $\forall x : \varphi_e(x) = e$.

There is a program that outputs its own length.

There is a program that outputs its own source code.

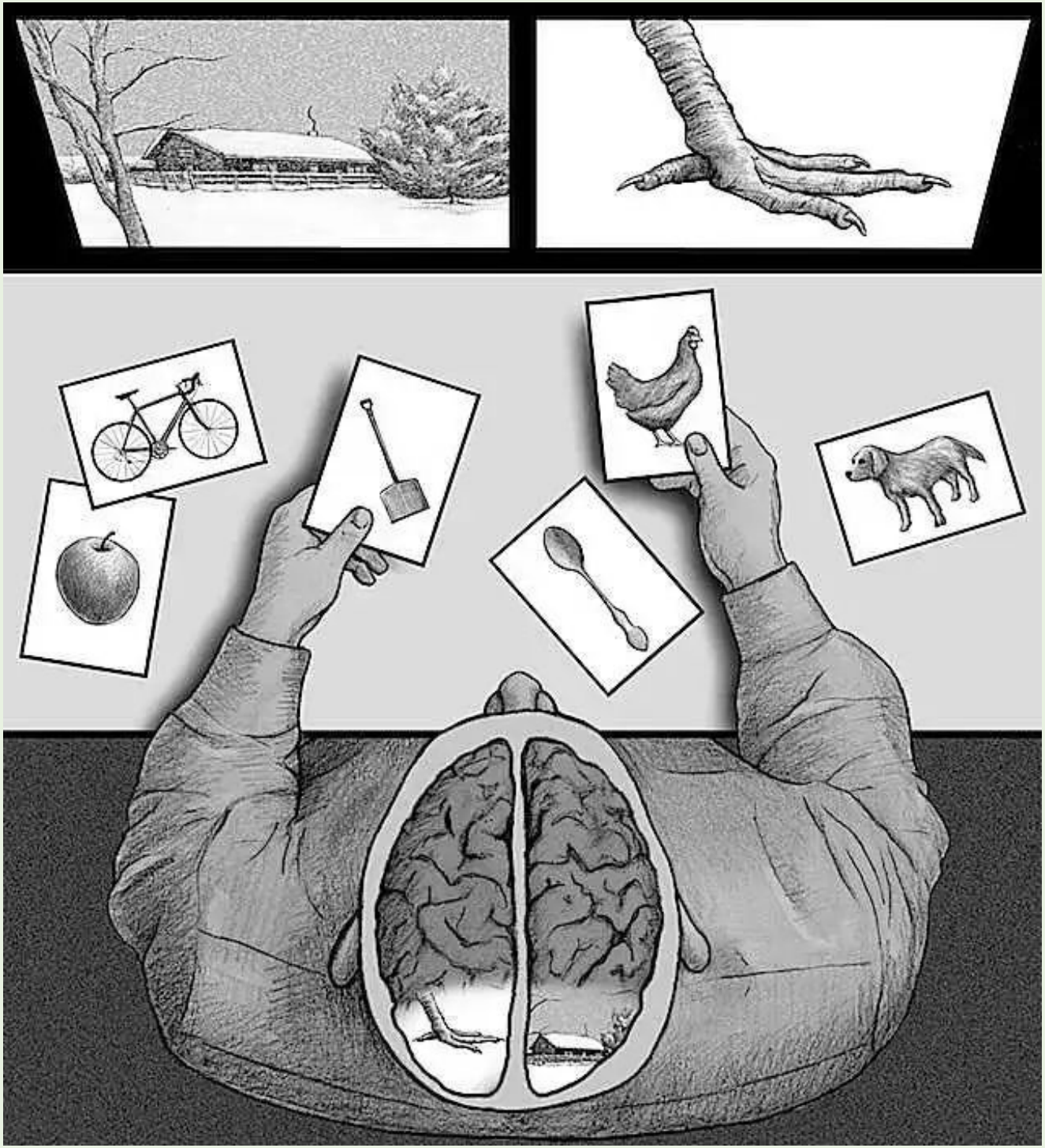
DNA / mutation / evolution

Introspective Program

There is a program that is totally introspective.

$$\varphi_e = \varphi_{h(e)}$$

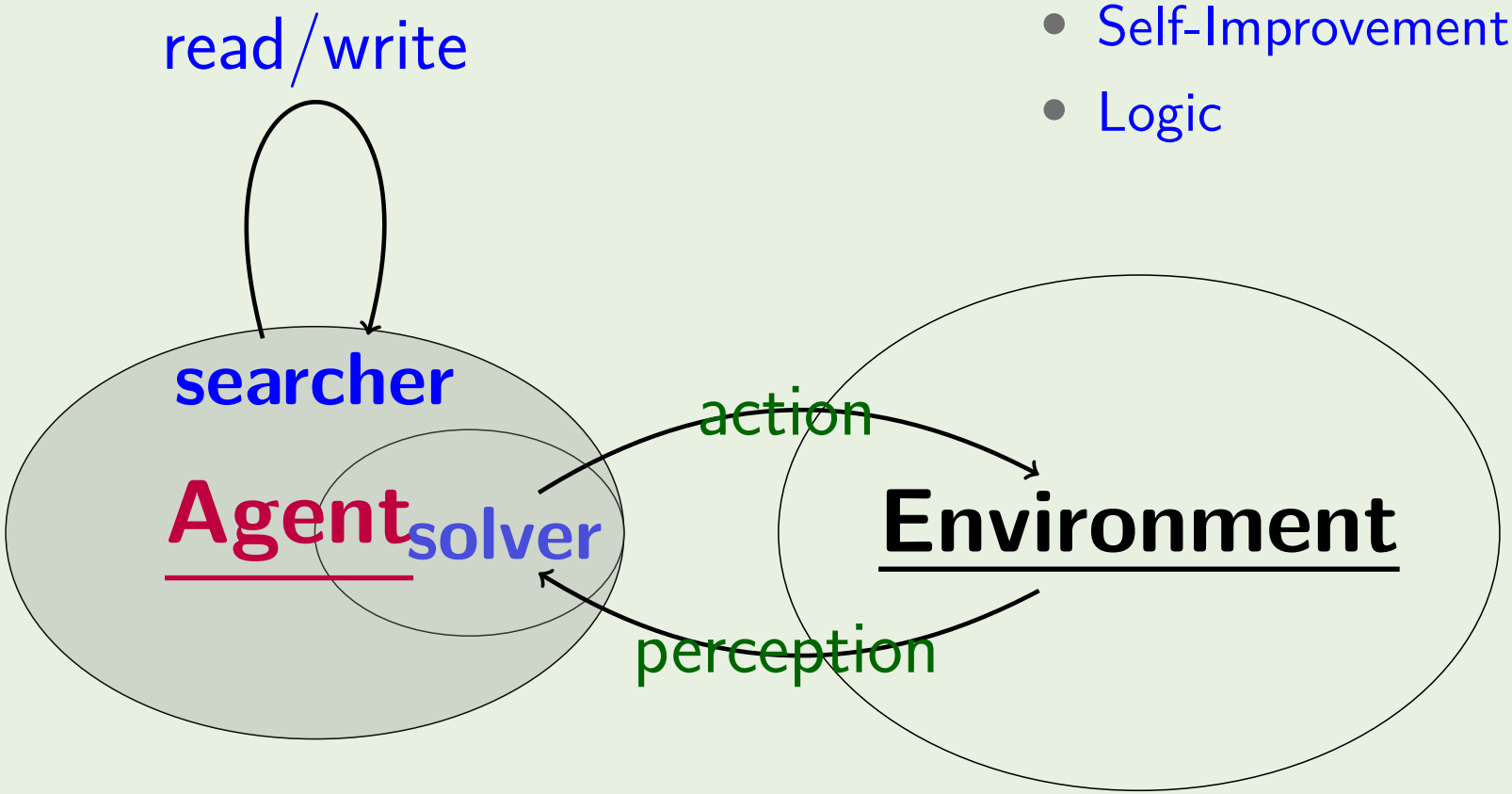
Self-simulating Computer	Self-consciousness
Host Machine	Experiencing Self
Virtual Machine	Remembering Self
Hardware	Body



说谎者悖论	**我在说谎**。
Grelling 悖论	‘非自谓的’是自谓的吗？
Russell 悖论	“不属于自身的集合的集合”属于自身吗？
Berry 悖论	我是少于十八个字不可定义的最小数。
Yablo 悖论	我下一句及后面所有的句子都是假的。
Gödel 不动点引理	**我有性质** α。
Tarski 算术真不可定义定理	我不真。
Gödel 第一不完全性定理	我不可证。
Gödel-Rosser 不完全性定理	对于任何一个关于我的证明，都有一个更短的关于我的否定的证明。
Löb 定理	如果我可证，那么 φ。
Curry 悖论	如果我是真的，那么圣诞老人存在。
Parikh 定理	我没有关于自己的长度短于 n 的证明。
Kleene 不动点定理	**我要进行** h **操作**。
Quine 悖论	把“把中的第一个字放到左引号前面，其余的字放到右引号后面，并保持引号及其中的字不变得到的句子是假的。”中的第一个字放到左引号前面，其余的字放到右引号后面，并保持引号及其中的字不变得到的句子是假的。
自测量长度程序	我要输出自己的长度。
自复制程序	我要输出自己。
自反省程序	我要回顾自己走过的每一步。
Gödel 机	我要变成能获取更大效用的自己。

Gödel Machine

- GRL
- Universal Search
- Self-Improvement
- Logic

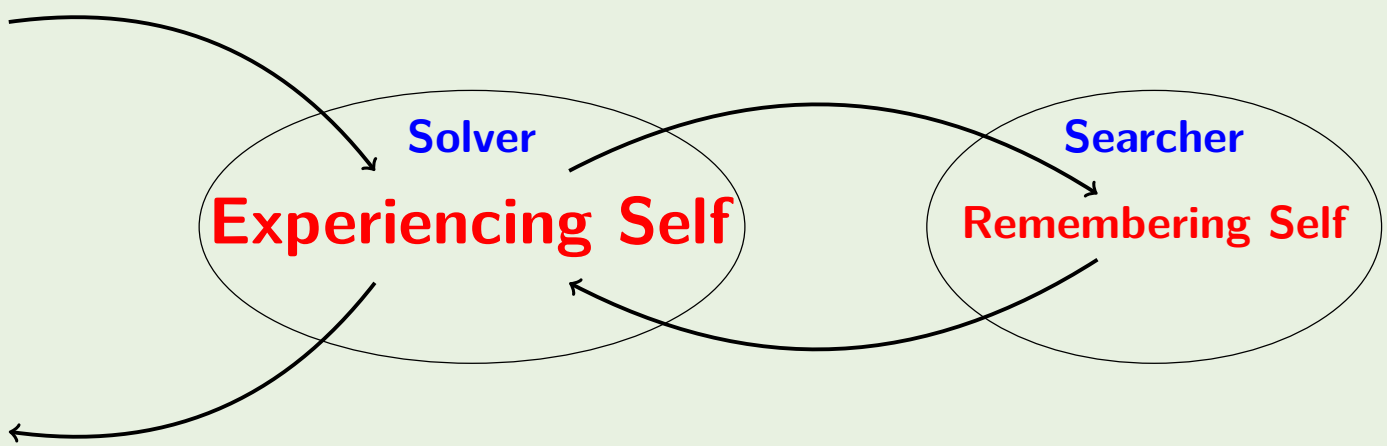


Disadvantage: A Gödel Machine with a badly chosen utility function is motivated to converge to a "poor" program. (orthogonality!)

Gödel Machine vs Self-Consciousness vs Free Will?

Self-simulating Computer	Gödel Machine	Self-consciousness
Host Machine	Solver	Experiencing Self
Virtual Machine	Searcher	Remembering Self
Hardware	Hardware	Body

$$\varphi_e = \varphi_{h(e)}$$



self-reference $\stackrel{?}{\Longrightarrow}$ self-improvement

Gödel Machines

1. *one-time* self-improvement: Kleene's fixpoint theorem

$$\varphi_e = \varphi_{h(e)}$$

2. *continuous* self-improvement: Kleene's fixpoint theorem with parameters

$$\varphi_{e(y)} = \varphi_{h(e(y),y)}$$

3. *uncomputable* case: Kleene's relativized fixpoint theorem

$$\varphi^A_{e(y)} = \varphi^A_{h(e(y),y)}$$

Limitation

1. Gödel's first incompleteness theorem / Rice's theorem
2. Gödel's second incompleteness theorem

$$\mathbb{T} \vdash \Box_{\mathbb{T}'} \varphi \to \varphi \implies \mathbb{T} \vdash Con(\mathbb{T}')$$

3. Legg's incompleteness theorem. *General prediction algorithms must be complex. Beyond a certain complexity they can't be mathematically discovered.*
4. Complexity: higher-level abstractions — coarse grained.
   Learning is to forget!


Thanks