

Self-reference, Computer Science, and Logic

Thomas Bolander,
Ph.D. student at IMM

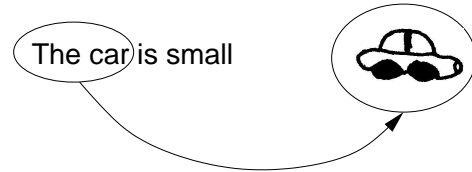
- *Self-reference* is in particular famous for the central role it plays in the classical paradoxes.
- As such, self-reference might not be considered as something to be taken serious.
- But, as we will see, self-reference and the arguments used in the paradoxes play a central role in many important results showing the *limitations* of computations, logic, and introspection in AI.

1

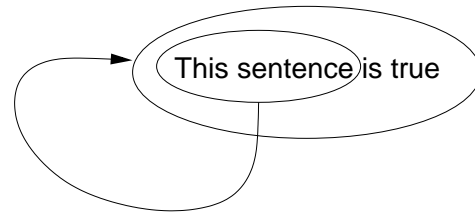
Self-Reference

Self-reference is used to denote any situation in which someone or something refers to itself.

Example. A non-self-referential sentence:



and a self-referential sentence:



2

Paradoxes

- The problem with self-reference: *paradoxes*.
- A **paradox** is a seemingly sound piece of reasoning based on seemingly true assumptions, that leads to a contradiction.

3

Grelling's Paradox

- A predicate is called **homological** if it is true of itself. Examples:
an abstract concept written upside down
- If a predicate is *not* true of itself it is called **heterological**. Example:
elephant

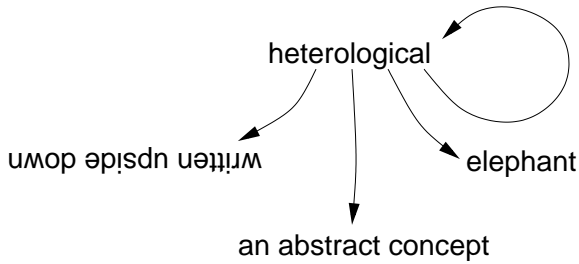
Now: *is "heterological" heterological?*

Assume the answer is "yes": then "heterological" is not true of itself. Therefore it is *not* heterological. But that means the answer is "no".

Assume the answer is "no": then "heterological" is true of itself. Therefore it *is* heterological. But that means the answer is "yes".

4

- Grelling's Paradox is *self-referential*: the definition of the predicate "heterological" refers to *all* predicates, including the predicate "heterological" itself.



- In particular, the sentence

Is "heterological" heterological?

is self-referential.

5

- A **formal system** consists of *formulas* and *proofs*. A formula can contain *free variables*:

$$p(x) \vee q(x),$$

which can be *instantiated* with numbers:
 $p(42) \vee q(42)$.

- The formulas of a formal system can be enumerated: $\varphi_1, \varphi_2, \varphi_3, \dots$
- For any formula ψ , $\lceil \psi \rceil$ is used to denote the i such that $\psi = \varphi_i$.
- A **proof** is a sequence of formulas $\varphi_{i_1}, \dots, \varphi_{i_n}$ satisfying certain requirements.
- For every formal system there should be a computer program which can decide whether a sequence of formulas $\varphi_{i_1}, \dots, \varphi_{i_n}$ is a proof in the system or not.

6

Computer Programs

- A **computer program** is something which is given an *input*, and that possibly returns an *output*.
- A computer program P is said to **accept** input s if it returns output "yes" on input s . If it returns output "no" on input s it is said to **reject** input s .



- The set of computer programs can be enumerated:

$$P_1, P_2, P_3, \dots$$

7

Formalizing the Paradox

The argument in Grelling's Paradox can be used *constructively* (non-paradoxically) to prove important properties of computer programs and formal systems.

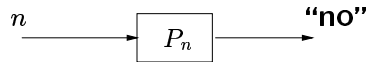
We will use formalized versions of Grelling's Paradox to prove:

- (a) There are sets M for which no computer program exists, that accepts as input exactly the elements of M .
- (b) In every sufficiently strong formal system there are formulas that can neither be proved nor disproved.
- (c) Logically based AI robots can not consistently believe themselves to have complete introspection.

8

The Halting Problem

A computer program P_n is called a **heterological program** if it rejects input n . In this case n is called a **heterological number**.



Theorem 1. *There exists no computer program P_h that accepts exactly the heterological numbers.*

Proof. Assume P_h exists, that is, assume

P_h accepts input $n \Leftrightarrow n$ is a het. number.

By definition of *heterological number* we have

n is a het. number $\Leftrightarrow P_n$ rejects input n .

These two equivalences together gives us

P_h accepts input $n \Leftrightarrow P_n$ rejects input n .

Letting $n = h$, we in particular get

P_h accepts input $h \Leftrightarrow P_h$ rejects input h .

This is a contradiction. □

9

That the proof above is nothing more than a variant of Grelling's Paradox is seen by the following correspondences:

<i>In Grelling's Paradox:</i>	<i>In the proof:</i>
predicate "heterological"	program P_h
is "het." heterological?	is P_h a het. prog.?

Asking whether P_h is a heterological program is the same as asking whether h is a heterological number.

10

Gödel's Incompleteness Theorem

Theorem 2. *In any consistent formal system containing a formula $\varphi(x)$ satisfying*

$\vdash \varphi(n) \Leftrightarrow n$ is a heterological number,

there will be a formula ψ such that neither ψ nor $\neg\psi$ can be proved.

Proof. Assume every formula can either be proved or disproved. By assumption there is a computer program P which can decide whether something is a proof. From P we can construct another program Q that given input n works like this:

Feed sequences of formulas successively into P until a sequence is found which is a proof of either $\varphi(n)$ or $\neg\varphi(n)$. If a proof of $\varphi(n)$ is found, return "yes", otherwise "no".

Q accepts exactly the heterological numbers, but this is proven to be impossible. □

11

Agent Introspection

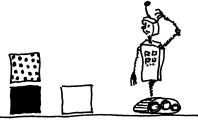
Agent: An AI system, e.g. a robot.

Introspective agent: Agent that can reflect on itself, its own thoughts and beliefs.

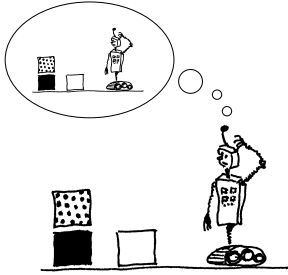
12

An Agent in Blocks World

Example. The world:



The agent contains a model of the world which it uses for planning and reasoning:

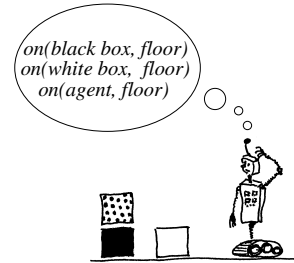


13

An agent's model of the world is often represented as a set of logical formulas such as:

$on(black\ box, floor)$
 $on(dotted\ box, black\ box)$
 $on(white\ box, floor)$
 $on(agent, floor)$.

These formulas are the agent's **beliefs** about the world.

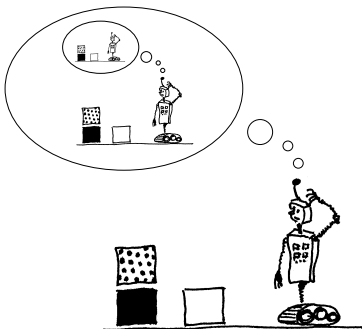


The agent's model of its world is the formal system in which these formulas are taken as axioms. The agent believes a formula φ iff $\vdash \varphi$ in this system.

14

Introspective Agent

If the agent is *introspective* its model of the world contains a model of the agent itself:



This corresponds to the agent having beliefs about its own beliefs.

If the agent believes the sentence

$on(black\ box, floor)$

to be among its own beliefs, that could be represented e.g. by

$I\ believe(\ulcorner on(black\ box, floor) \urcorner)$.

15

Complete Introspection

An agent is said to have **complete introspection** if its beliefs about itself are correct and complete.

Formally, this means that an agent has complete introspection if:

$\vdash \varphi \Leftrightarrow \vdash I\ believe(\ulcorner \varphi \urcorner)$

for all φ .

The agent might not have complete introspection. But it might still *believe* itself to have complete introspection. That would correspond to having:

$\vdash \varphi \leftrightarrow I\ believe(\ulcorner \varphi \urcorner)$ for all φ .

16

Theorem 3. Any agent that believes itself to have complete introspection will have contradictory beliefs.

Proof. We assume that we are given a formal system in which

$$\vdash \varphi \leftrightarrow I \text{ believe}(\ulcorner \varphi \urcorner) \quad \text{for all } \varphi. \quad (1)$$

We say that a formula $\varphi(x)$ is a **heterological formula** if $\neg I \text{ believe}(\ulcorner \varphi(\ulcorner \varphi \urcorner) \urcorner)$ is provable. For every formula φ we define

$$\text{het}(\varphi) =_{df} \neg I \text{ believe}(\ulcorner \varphi(\ulcorner \varphi \urcorner) \urcorner).$$

Now we ask whether $\text{het}(\ulcorner \text{het}(x) \urcorner)$ holds or not (cf. Grelling's Paradox). By definition of *het* we have

$$\vdash \text{het}(\ulcorner \text{het}(x) \urcorner) \leftrightarrow \neg I \text{ believe}(\ulcorner \text{het}(\ulcorner \text{het}(x) \urcorner) \urcorner).$$

As an instance of (1) we get

$$\vdash \text{het}(\ulcorner \text{het}(x) \urcorner) \leftrightarrow I \text{ believe}(\ulcorner \text{het}(\ulcorner \text{het}(x) \urcorner) \urcorner)$$

These two equivalences taken together shows the formal system to be inconsistent. \square

- Self-reference is something to be taken serious when talking about computations, logic, and artificial intelligence.
- Self-reference shows that a lot of things are impossible, that we perhaps would have expected to be possible.
- If we want to circumvent the impossibility results we should try to get to a deeper theoretical understanding of the nature of self-reference.