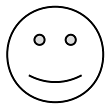
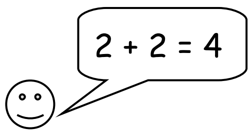
(The Cartoon Guide to) Löb's Theorem

by Eliezer Yudkowsky (http://yudkowsky.net)

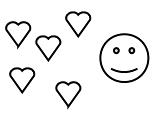
with thanks to Torkel Franzén and Marcello Herreshoff



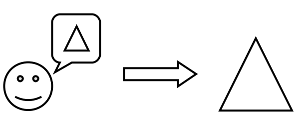
This is our friend PA. His full name is "Peano Arithmetic."



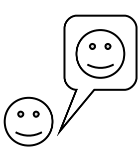
PA makes statements about mysterious mathematical objects called "numbers".



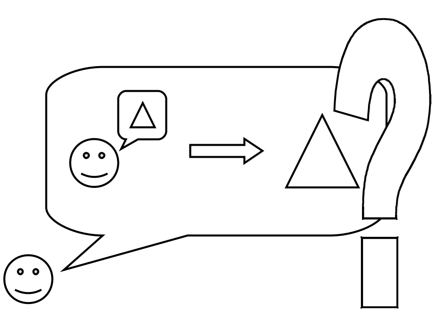
We all trust our PA. When PA tells us something, we call it a *proof.*



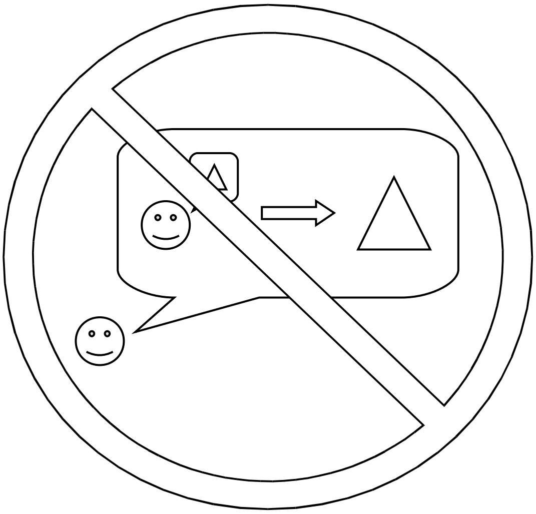
In our experience so far, PA is *sound* - if PA says something, it always turns out to be true. So we generalize the rule "PA is always right", by scientific induction.



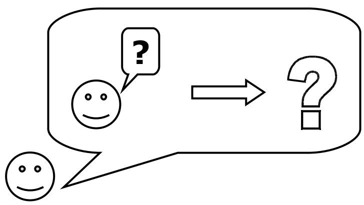
A clever fellow named Gödel figured out how to ask PA questions about PA.



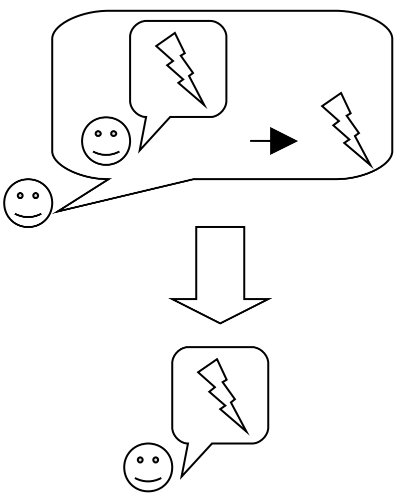
Does PA claim to be sound? Then we could trust PA even more - since, in our experience so far, scientific induction is sometimes wrong, but PA is always right.



But that darned Gödel proved this impossible - if PA asserted its own soundness (in general), PA would become inconsistent.



Maybe we could prove soundness (in PA) for some special cases? But if so, which statements could we prove sound?



Martin Hugo Löb proved a surprising theorem: If PA proves "If PA proves '**X**', then **X**", then PA proves **X**!

That is: PA can prove that a proof of '**X**' implies **X**, only if PA can prove **X**. So if we found a proof (within PA) that "If Peano Arithmetic proves Goldbach's Conjecture, it is correct", we could use this fact to prove Goldbach's Conjecture!

Alas, this means we can't prove PA sound with respect to any important class of statements.

For example, we want to prove "If PA proves 'X + Y = Z', then X + Y = Z".

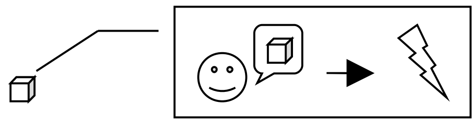
"If PA proves '1 + 2 = 5', then 1 + 2 = 5" is a statement in that class.

The hope would be to show that if 1 + 2 ≠ 5, PA does *not* prove "1 + 2 = 5".

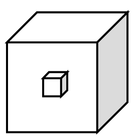
Unfortunately, Löb's Theorem demonstrates that if we could prove the above *within* PA, then PA would prove 1 + 2 = 5.

Löb's Sentence:

The key to Löb's Theorem is the Löb sentence **L**.
L states: "If a proof of **L** exists, then ***C***".
C is some arbitrary statement like "1 + 2 = 5".
We don't start with a *proof* of **L** - we just construct the *sentence.*

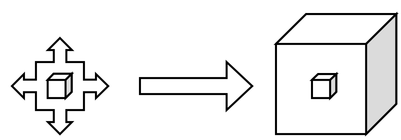


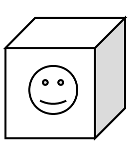
A paradoxical analogue of **L** is the *Santa Claus sentence:* "If this sentence is true, then Santa Claus exists." (Clearly, if that sentence were really true, Santa Claus would have to exist. But this is just what the sentence asserts, so it is true, and Santa Claus does exist.) **L** itself is not paradoxical, because it talks about *proof* within a system, not *truth*.



Direct self-reference is not permitted in Peano Arithmetic. Yet the Löb sentence **L** apparently refers to **L** itself.

For a sentence to refer to itself, it must contain a self-replicating recipe - when the recipe is executed, it produces a copy of the complete sentence, including the recipe itself.

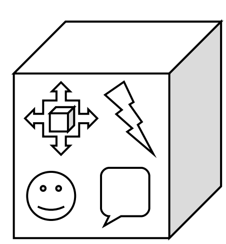




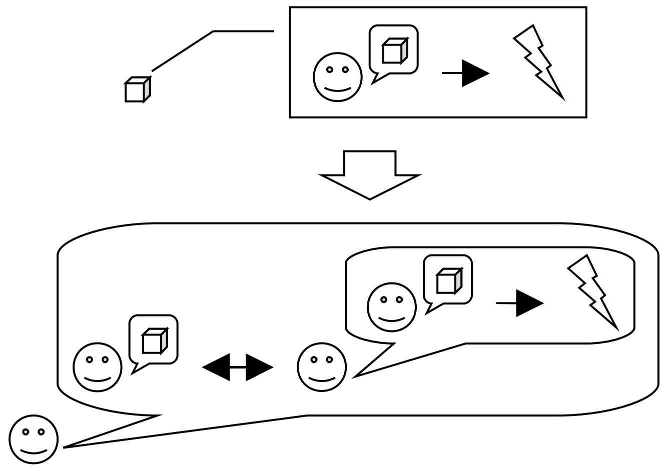
Likewise, Peano Arithmetic does not contain any direct method of referring to "provability". For an arithmetical sentence to talk about "proof", it must contain a (huge) recipe for the formal system of Peano Arithmetic - a numerical encoding of PA's axioms and inference rules.

A non-paradoxical version of **L** in English:

If the result of substituting "If the result of substituting x for 'x' in x is provable in PA, then C" for 'x' in "If the result of substituting x for 'x' in x is provable in PA, then C" is provable in PA, then C.



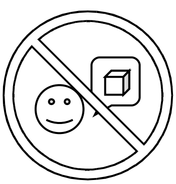
Our Plan:



Step 1: Unpack the box.

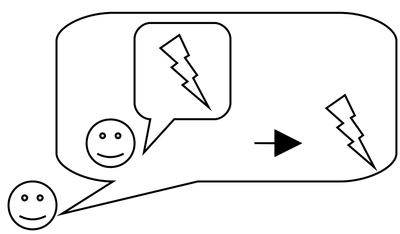
We first prove (within PA) that "PA proves **L** if and only if PA proves that a proof of **L** implies C."

Easy: this follows trivially from the construction of **L**.



Note that we have *not* proved **L** itself, just that PA proves **L** (the box) if and only if PA proves "Proving **L** implies ***C***" (the content of the box). We can always *construct* the sentence **L**, but that doesn't mean we can *prove* it.

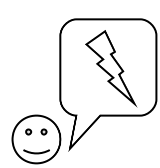
Step 2: The Löb hypothesis.



The Löb hypothesis is that PA proves "A proof of ***C*** implies ***C***." For example, ***C*** could be "1 + 2 = 5", in which case Löb's hypothesis is that PA proves "If PA proves '1 + 2 = 5', then 1 + 2 = 5." Observe how the Löb *hypothesis* differs from the Löb *sentence*.

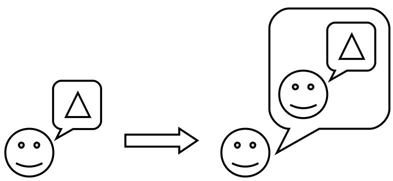
Step 3: Profit.

Starting from these two steps, we will prove ***C*** within PA.

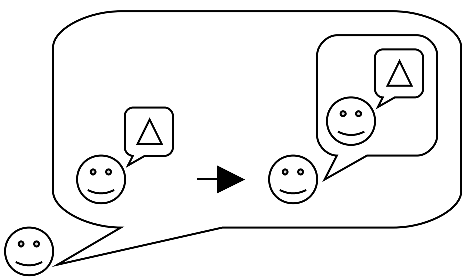


Ingredients:

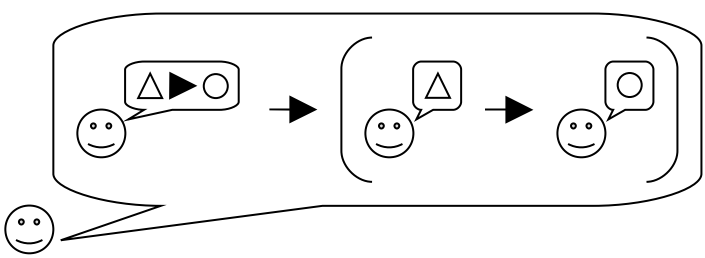
The formal proof uses a few more ingredients. These are not additional hypotheses; they are universal properties of Peano Arithmetic, and provable within PA itself.



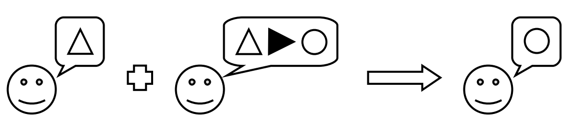
A1: If PA proves **X**, PA proves that "PA proves **X**".



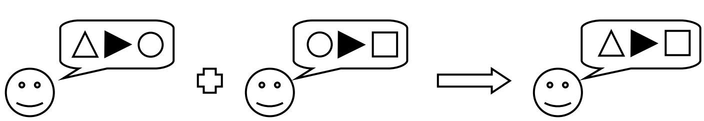
A2: PA can prove A1.



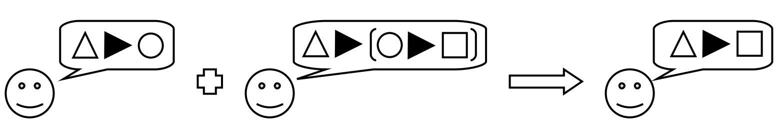
A3: PA proves that PA obeys Modus Ponens.



MP: Modus Ponens

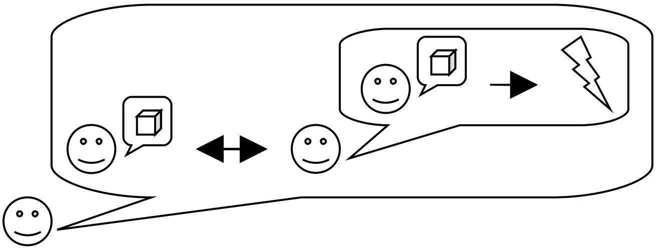


B1: (A->B) + (B->C) => (A->C)

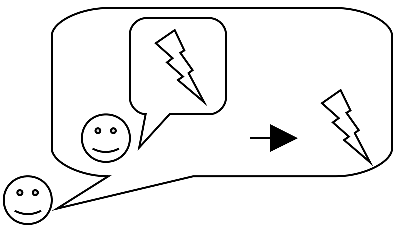


B2: (A->B) + A->(B->C) => (A->C)

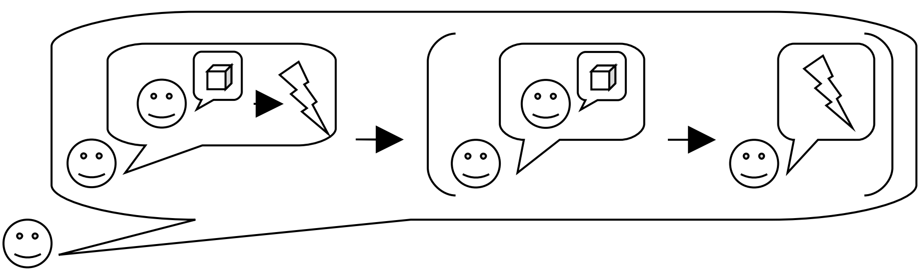
The Proof:



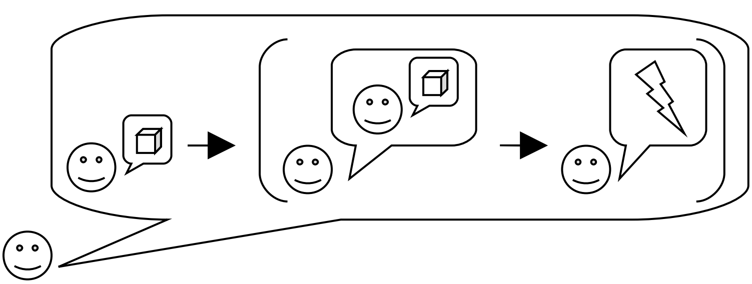
1. Unpack the box.



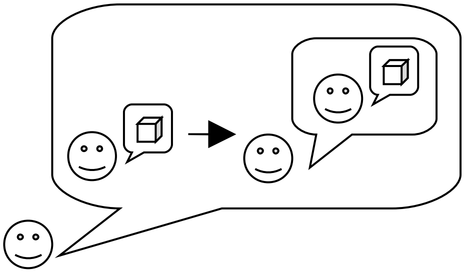
2. Löb's hypothesis.



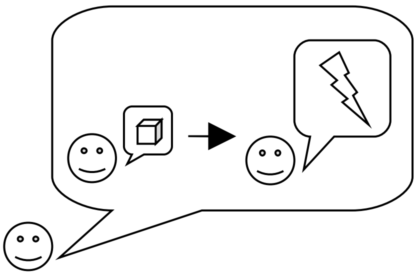
3. (A3)



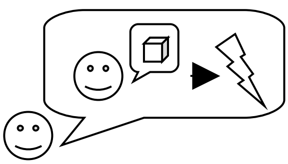
4. (1, 3, B1)



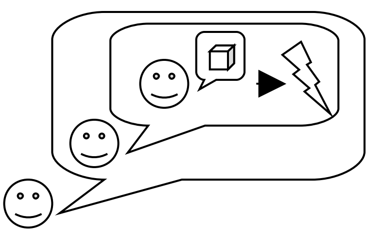
5.
(A2)



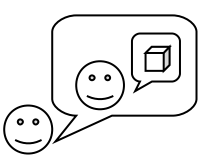
6.
(5, 4, B2)



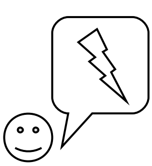
7.
(6, 2, B1)



8.
(7, A1)



9.
(1, MP)



10.
(9, 7, MP)

Q.E.D.