# Löb's Theorem

## A functional pearl of dependently typed quining

Anonymous

***Keywords*** Agda, Löb's theorem, quine, self-reference, type theory

## Abstract

Löb's theorem states that to prove that a proposition is provable, it is sufficient to prove the proposition under the assumption that it is provable. The Curry-Howard isomorphism identifies formal proofs with abstract syntax trees of programs; Löb's theorem thus states that self-interpreters are impossible for total languages. We formalize a few variations of Löb's theorem in Agda using an inductive-inductive encoding of terms indexed over types. We verify the consistency of our formalizations relative to Agda by giving them semantics via interpretation functions.

> *If P's answer is 'Bad!', Q will suddenly stop.*
> *But otherwise, Q will go back to the top,*
> *and start off again, looping endlessly back,*
> *till the universe dies and turns frozen and black.*
>
> Excerpt from *Scooping the Loop Snooper: A proof that the Halting Problem is undecidable* (Pullum 2000))

## 1. Introduction

Löb's theorem has a variety of applications, from providing an induction rule for program semantics involving a "later" operator (Appel et al. 2007), to proving incompleteness of a logical theory as a trivial corollary, from acting as a no-go theorem for a large class of self-interpreters, to allowing robust cooperation in the Prisoner's Dilemma with Source Code (Barasz et al. 2014), and even in one case curing social anxiety (Yudkowsky 2014).

In this paper, after introducing the content of Löb's theorem, we will present three formalizations in Agda: one that shows the theorem is admissible as an axiom in a wide range of situations, one which prove's Löb's theorem by assuming quines, and one which constructs the proof under even weaker assumptions; see section 5 for details.

"What is Löb's theorem, this versatile tool with wondrous applications?" you may ask.

Consider the sentence "if this sentence is true, then you, dear reader, are the most awesome person in the world." Suppose that this sentence is true. Then you, dear reader are the most awesome person in the world. Since this is exactly what the sentence asserts, the sentence is true, and you, dear reader, are the most awesome person in the world. For those more comfortable with symbolic logic, we can let $X$ be the statement "you, dear reader, are the most awesome person in the world", and we can let $A$ be the statement "if this sentence is true, then $X$". Since we have that $A$ and $A \to X$ are the same, if we assume $A$, we are also assuming $A \to X$, and hence we have $X$. Thus since assuming $A$ yields $X$, we have that $A \to X$. What went wrong?[1]

It can be made quite clear that something is wrong; the more common form of this sentence is used to prove the existence of Santa Claus to logical children: considering the sentence "if this sentence is true, then Santa Claus exists", we can prove that Santa Claus exists. By the same logic, though, we can prove that Santa Claus does not exist by considering the sentence "if this sentence is true, then Santa Claus does not exist." Whether you consider it absurd that Santa Claus exist, or absurd that Santa Claus not exist, surely you will consider it absurd that Santa Claus both exist and not exist. This is known as Curry's Paradox.

Have you figured out what went wrong?

The sentence that we have been considering is not a valid mathematical sentence. Ask yourself what makes it invalid, while we consider a similar sentence that is actually valid.

Now consider the sentence"if this sentence is provable, then you, dear reader, are the most awesome person in the world." Fix a particular formalization of provability (for example, Peano Arithmetic, or Martin–Löf Type Theory). To prove that this sentence is true, suppose that it is provable. We must now show that you, dear reader, are the most awesome person in the world. *If provability implies truth*, then the sentence is true, and then you, dear reader, are the most awesome person in the world. Thus, if we can assume that provability implies truth, then we can prove that the sentence is true. This, in a nutshell, is Löb's theorem: to prove $X$, it suffices to prove that $X$ is true whenever $X$ is provable. Symbolically, this is

$$\Box(\Box X \to X) \to \Box X$$

where $\Box X$ means "$X$ is provable" (in our fixed formalization of provability).

Let us now return to the question we posed above: what went wrong with our original sentence? The answer is that self-reference with truth is impossible, and the clearest way we know to argue for this is via the Curry–Howard Isomorphism; in a particular technical sense, the problem is that self-reference with truth fails to terminate.

The Curry–Howard Isomorphism establishes an equivalence between types and propositions, between (well-typed, terminating, functional) programs and proofs. See Table 1 for some examples. Now we ask: what corresponds to a formalization of provability?

[1] Those unfamiliar with conditionals should note that the "if … then …" we use here is the logical "if", where "if false then $X$" is always true, and not the counter-factual "if".

| Logic | Programming | Set Theory |
|---|---|---|
| Proposition | Type | Set of Proofs |
| Proof | Program | Element |
| Implication ($\to$) | Function ($\to$) | Function |
| Conjunction ($\wedge$) | Pairing (,) | Cartesian Product ($\times$) |
| Disjunction ($\vee$) | Sum (+) | Disjoint Union ($\sqcup$) |
| Gödel codes | ASTs | — |

**Table 1.** The Curry-Howard isomorphism between mathematical logic and functional programming

If a proof of P is a terminating functional program which is well-typed at the type corresponding to P, and to assert that P is provable is to assert that the type corresponding to P is inhabited, then an encoding of a proof is an encoding of a program. Although mathematicians typically use Gödel codes to encode propositions and proofs, a more natural choice of encoding programs will be abstract syntax trees. In particular, a valid syntactic proof of a given (syntactic) proposition corresponds to a well-typed syntax tree for an inhabitant of the corresponding syntactic type.

Unless otherwise specified, we will henceforth consider only well-typed, terminating programs; when we say "program", the adjectives "well-typed" and "terminating" are implied.

## 2. Quines

What is the computational equivalent of the sentence "If this sentence is provable, then $X$"? It will be something of the form "??? $\to X$". As a warm-up, let's look at a Python program that returns a string representation of this type.

To do this, we need a program that outputs its own source code. There are three genuinely distinct solutions, the first of which is degenerate, and the second of which is cheeky (or sassy?). These "cheating" solutions are:

- The empty program, which outputs nothing.
- The code `print(open(__file__, 'r').read())`, which relies on the Python interpreter to get the source code of the program.

Now we develop the standard solution. At a first gloss, it looks like:

```
(lambda T: '(' + T + ') → X') "???"
```

Now we need to replace `"???"` with the entirety of this program code. We use Python's string escaping function (`repr`) and replacement syntax (`("foo %s bar" % "baz")` becomes `"foo baz bar"`):

```
(lambda T: '(' + T % repr(T) + ') → X')
 ("(lambda T: '(' + T %% repr(T) + ') → X')\n (%s)")
```

This is a slight modification on the standard way of programming a quine, a program that outputs its own source-code.

Suppose we have a function $\Box$ that takes in a string representation of a type, and returns the type of syntax trees of programs producing that type. Then our Löbian sentence would look something like (if $\to$ were valid notation for function types in Python)

```
(lambda T: □ (T % repr(T)) → X)
 ("(lambda T: □ (T %% repr(T)) → X)\n (%s)")
```

Now, finally, we can see what goes wrong when we consider using "if this sentence is true" rather than "if this sentence is provable". Provability corresponds to syntax trees for programs; truth corresponds to execution of the program itself. Our pseudo-Python thus becomes

```
(lambda T: eval(T % repr(T)) → X)
 ("(lambda T: eval(T %% repr(T)) → X)\n (%s)")
```

This code never terminates! So, in a quite literal sense, the issue with our original sentence was that, if we tried to phrase it, we'd never finish.

Note well that the type ($\Box$ X $\to$ X) is a type that takes syntax trees and evaluates them; it is the type of an interpreter.

## 3. Abstract Syntax Trees for Dependent Type Theory

The idea of formalizing a type of syntax trees which only permits well-typed programs is common in the literature (McBride 2010; Chapman 2009; Danielsson 2007). For example, here is a very simple (and incomplete) formalization with dependent function types ($\Pi$), a unit type ($\top$), an empty type ($\bot$), and functions ($\lambda$).

We will use some standard data type declarations, which are provided for completeness in Appendix A.

```
mutual
  infixl 2 _▷_

  data Context : Set where
    ε : Context
    _▷_ : (Γ : Context) → Type Γ → Context

  data Type : Context → Set where
    `⊤ : ∀ {Γ} → Type Γ
    `⊥ : ∀ {Γ} → Type Γ
    `Π : ∀ {Γ}
      → (A : Type Γ) → Type (Γ ▷ A) → Type Γ

  data Term : {Γ : Context} → Type Γ → Set where
    `tt : ∀ {Γ} → Term {Γ} `⊤
    `λ : ∀ {Γ A B} → Term B → Term {Γ} (`Π A B)
```

An easy way to check consistency of a syntactic theory which is weaker than the theory of the ambient proof assistant is to define an interpretation function, also commonly known as an unquoter, or a denotation function, from the syntax into the universe of types. Here is an example of such a function:

```
mutual
  ⟦_⟧ᶜ : Context → Set
  ⟦ ε ⟧ᶜ = ⊤
  ⟦ Γ ▷ T ⟧ᶜ = Σ ⟦ Γ ⟧ᶜ ⟦ T ⟧ᵀ

  ⟦_⟧ᵀ : ∀ {Γ}
    → Type Γ → ⟦ Γ ⟧ᶜ → Set
  ⟦ `⊤ ⟧ᵀ Γ⇓ = ⊤
  ⟦ `⊥ ⟧ᵀ Γ⇓ = ⊥
  ⟦ `Π A B ⟧ᵀ Γ⇓ = (x : ⟦ A ⟧ᵀ Γ⇓) → ⟦ B ⟧ᵀ (Γ⇓ , x)

  ⟦_⟧ᵗ : ∀ {Γ T}
    → Term {Γ} T → (Γ⇓ : ⟦ Γ ⟧ᶜ) → ⟦ T ⟧ᵀ Γ⇓
  ⟦ `tt ⟧ᵗ Γ⇓ = tt
  ⟦ `λ f ⟧ᵗ Γ⇓ x = ⟦ f ⟧ᵗ (Γ⇓ , x)
```

Note that this interpretation function has an essential property that we will call *locality*: the interpretation of any given constructor does not require doing case analysis on any of its arguments. By contrast, one could imagine an interpretation function that interpreted function types differently depending on their domain and codomain; for example, one might interpret `⊥ `→ A as $\top$.

## 4. This Paper

In this paper, we make extensive use of this trick for validating models. In section 6, we formalize the simplest syntax that supports Löb's theorem and prove it sound relative to Agda in 12 lines of code; the understanding is that this syntax could be extended to support basically anything you might want. We then present in section 7 an extended version of this solution, which supports enough operations that we can prove our syntax sound (consistent), incomplete, and nonempty. In a hundred lines of code, we prove Löb's theorem in section 8 under the assumption that we are given a quine; this is basically the well-typed functional version of the program that uses `open(__file__, 'r').read()`. After taking a digression for an application of Löb's theorem to the prisoner's dilemma in section 9, we sketch in section 10 our implementation of Löb's theorem (code in the supplemental material) based on only the assumption that we can add a level of quotation to our syntax tree; this is the equivalent of letting the compiler implement `repr`, rather than implementing it ourselves. We close in section 11 with some discussion about avenues for removing the hard-coded `repr`.

## 5. Prior Work

There exist a number of implementations or formalizations of various flavors of Löb's theorem in the literature. Appel et al. use Löb's theorem as an induction rule for program logics in Coq (Appel et al. 2007). Piponi formalizes a rule with the same shape as Löb's theorem in Haskell, and uses it for, among other things, spreadsheet evaluation (Piponi 2006).

Gödel's incompleteness theorems, easy corrolaries to Löb's theorem, have been formally verified numerous times (Shankar 1986, 1997; O'Connor 2005).

To our knowledge, our twelve line proof is the shortest self-contained formally verified proof of the admissibility of Löb's theorem to date. We are not aware of other formally verified proofs of Löb's theorem which interpret the modal $\Box$ operator as an inductively defined type of syntax trees of proofs of a given theorem, as we do in this formalization. Finally, we are not aware of other work which uses the trick of talking about a local interpretation function (as described at the end of section 3) to talk about consistent extensions to classes of encodings of type theory.

## 6. Trivial Encoding

We begin with a language that supports almost nothing other than Löb's theorem. We use $\Box$ `T` to denote the type of `Terms` of whose syntactic type is `T`. We use `‘□’ T` to denote the syntactic type corresponding to the type of (syntactic) terms whose syntactic type is `T`. For example, the type of a `repr` which operated on syntax trees would be $\Box$ `T` $\to$ $\Box$ `(‘□’ T)`.

```
data Type : Set where
  _‘→’_ : Type → Type → Type
  ‘□’ : Type → Type

data □ : Type → Set where
  Löb : ∀ {X} → □ (‘□’ X ‘→’ X) → □ X
```

The only term supported by our term language is Löb's theorem. We can prove this language consistent relative to Agda with an interpreter:

```
⟦_⟧ᵀ : Type → Set
⟦ A ‘→’ B ⟧ᵀ = ⟦ A ⟧ᵀ → ⟦ B ⟧ᵀ
⟦ ‘□’ T ⟧ᵀ  = □ T

⟦_⟧ᵗ : ∀ {T : Type} → □ T → ⟦ T ⟧ᵀ
⟦ Löb □‘X’→X ⟧ᵗ = ⟦ □‘X’→X ⟧ᵗ (Löb □‘X’→X)
```

To interpret Löb's theorem applied to the syntax for a compiler $f$ (`□‘X’→X` in the code above), we interpret $f$, and then apply this interpretation to the constructor `Löb` applied to $f$.

Finally, we tie it all together:

```
löb : ∀ {‘X’} → □ (‘□’ ‘X’ ‘→’ ‘X’) → ⟦ ‘X’ ⟧ᵀ
löb f = ⟦ Löb f ⟧ᵗ
```

This code is deceptively short, with all of the interesting work happening in the interpretation of `Löb`.

What have we actually proven, here? It may seem as though we've proven absolutely nothing, or it may seem as though we've proven that Löb's theorem always holds. Neither of these is the case. The latter is ruled out, for example, by the existence of an self-interpreter for $F_\omega$ (Brown and Palsberg 2016).[2]

We have proven the following. Suppose you have a formalization of type theory which has a syntax for types, and a syntax for terms indexed over those types. If there is a "local explanation" for the system being sound, i.e., an interpretation function where each rule does not need to know about the full list of constructors, then it is consistent to add a constructor for Löb's theorem to your syntax. This means that it is impossible to contradict Löb's theorem no matter what (consistent) constructors you add. We will see in the next section how this gives incompleteness, and discuss in later sections how to *prove Löb's theorem*, rather than simply proving that it is consistent to assume.

## 7. Encoding with Soundness, Incompleteness, and Non-Emptiness

By augmenting our representation with top (`‘⊤’`) and bottom (`‘⊥’`) types, and a unique inhabitant of `‘⊤’`, we can prove soundness, incompleteness, and non-emptiness.

```
data Type : Set where
  _‘→’_ : Type → Type → Type
  ‘□’ : Type → Type
  ‘⊤’ : Type
  ‘⊥’ : Type

-- "□" is sometimes written as "Term"
data □ : Type → Set where
  Löb : ∀ {X} → □ (‘□’ X ‘→’ X) → □ X
  ‘tt’ : □ ‘⊤’

⟦_⟧ᵀ : Type → Set
⟦ A ‘→’ B ⟧ᵀ = ⟦ A ⟧ᵀ → ⟦ B ⟧ᵀ
⟦ ‘□’ T ⟧ᵀ  = □ T
⟦ ‘⊤’ ⟧ᵀ    = ⊤
⟦ ‘⊥’ ⟧ᵀ    = ⊥

⟦_⟧ᵗ : ∀ {T : Type} → □ T → ⟦ T ⟧ᵀ
⟦ Löb □‘X’→X ⟧ᵗ = ⟦ □‘X’→X ⟧ᵗ (Löb □‘X’→X)
⟦ ‘tt’ ⟧ᵗ = tt

‘¬’_ : Type → Type
‘¬’ T = T ‘→’ ‘⊥’
```

[2] One may wonder how exactly the self-interpreter for $F_\omega$ does not contradict this theorem. In private conversations with Matt Brown, we found that the $F_\omega$ self-interpreter does not have a separate syntax for types, instead indexing its terms over types in the metalanguage. This means that the type of Löb's theorem becomes either `□ (□ X → X) → □ X`, which is not strictly positive, or `□ (X → X) → □ X`, which, on interpretation, must be filled with a general fixpoint operator. Such an operator is well-known to be inconsistent.

```
löb : ∀ {‘X’} → □ (‘□’ ‘X’ ‘→’ ‘X’) → ⟦ ‘X’ ⟧ᵀ
löb f = ⟦ Löb f ⟧ᵗ

-- There is no syntactic proof of absurdity
soundness : ¬ □ ‘⊥’
soundness x = ⟦ x ⟧ᵗ

-- but it would be absurd to have a syntactic
-- proof of that fact
incompleteness : ¬ □ (‘¬’ (‘□’ ‘⊥’))
incompleteness = löb

-- However, there are syntactic proofs of some
-- things (namely ⊤)
non-emptiness : □ ‘⊤’
non-emptiness = ‘tt’

-- There are no syntactic interpreters, things
-- which, at any type, evaluate code at that
-- type to produce its result.
no-interpreters : ¬ (∀ {‘X’} → □ (‘□’ ‘X’ ‘→’ ‘X’))
no-interpreters interp = löb (interp {‘⊥’})
```

What is this incompleteness theorem? Gödel's incompleteness theorem is typically interpred as "there exist true but unprovable statements." In intuitionistic logic, this is hardly surprising. A more accurate rendition of the theorem in Agda might be "there exist true but inadmissible statements," i.e., there are statements which are provable meta-theoretically, but which lead to (meta-theoretically–provable) inconsistency if assumed at the object level.

This may seem a bit esoteric, so let's back up a bit, and make it more concrete. Let's begin by banishing "truth". Sometimes it is useful to formalize a notion of provability. For example, you might want to show neither assuming $T$ nor assuming $\neg T$ yields a proof of contradiction. You cannot phrase this is $\neg T \wedge \neg\neg T$, for that is absurd. Instead, you want to say something like $(\neg\Box T) \wedge \neg\Box(\neg T)$, i.e., it would be absurd to have a proof object of either $T$ or of $\neg T$. After a while, you might find that meta-programming in this formal syntax is nice, and you might want it to be able to formalize every proof, so that you can do all of your solving reflectively. If you're like us, you might even want to reason about the reflective tactics themselves in a reflective manner; you'd want to be able to add levels of quotation to quoted things to talk about such tactics.

The incompleteness theorem, then, is this: your reflective system, no matter how powerful, cannot formalize every proof. For any fixed language of syntactic proofs which is powerful enough to represent itself, there will always be some valid proofs that you cannot reflect into your syntax. In particular, you might be able to prove that your syntax has no proofs of $\bot$ (by interpreting any such proof). But you'll be unable to quote that proof. This is what the incompleteness theorem stated above says. Incompleteness, fundamentally, is a result about the limitations of formalizing provability.

## 8. Encoding with Quines

We now weaken our assumptions further. Rather than assuming Löb's theorem, we instead assume only a type-level quine in our representation. Recall that a *quine* is a program that outputs some function of its own source code. A *type-level quine at* $\phi$ is program that outputs the result of evaluating the function $\phi$ on the abstract syntax tree of its own type. Letting `Quine` $\phi$ denote the constructor for a type-level quine at $\phi$, we have an isomorphism between `Quine` $\phi$ and $\phi$ $\ulcorner$ `Quine` $\phi$ $\urcorner^{\mathsf{T}}$, where $\ulcorner$ `Quine` $\phi$ $\urcorner^{\mathsf{T}}$ is the abstract syntax tree for the source code of `Quine` $\phi$. Note that we assume constructors for "adding a level of quotation", turning abstract syntax trees for programs of type $T$ into abstract syntax trees for abstract syntax trees for programs of type $T$; this corresponds to `repr`.

```
infixl 3 _‘’ₐ_
infixl 3 _w‘‘’’ₐ_
infixl 3 _‘’_
infixl 2 _▷_
infixr 2 _‘∘’_
infixr 1 _‘→’_
```

We begin with an encoding of contexts and types, repeating from above the constructors of '→', '□', '⊤', and '⊥'. We add to this a constructor for quines (`Quine`), and a constructor for syntax trees of types in the empty context ('Type$\varepsilon$'). Finally, rather than proving weakening and substitution as mutually recursive definitions, we take the easier but more verbose route of adding constructors that allow adding and substituting extra terms in the context. Note that '□' is now a function of the represented language, rather than a meta-level operator.

Note that we use the infix operator `_‘’_` to denote substitution.

```
mutual
  data Context : Set where
    ε : Context
    _▷_ : (Γ : Context) → Type Γ → Context

  data Type : Context → Set where
    _‘→’_ : ∀ {Γ} → Type Γ → Type Γ → Type Γ
    ‘⊤’ : ∀ {Γ} → Type Γ
    ‘⊥’ : ∀ {Γ} → Type Γ
    ‘Typeε’ : ∀ {Γ} → Type Γ
    ‘□’ : ∀ {Γ} → Type (Γ ▷ ‘Typeε’)
    Quine : Type (ε ▷ ‘Typeε’) → Type ε
    W : ∀ {Γ A}
      → Type Γ → Type (Γ ▷ A)
    W₁ : ∀ {Γ A B}
      → Type (Γ ▷ B) → Type (Γ ▷ A ▷ (W B))
    _‘’_ : ∀ {Γ A}
      → Type (Γ ▷ A) → Term A → Type Γ
```

In addition to '$\lambda$' and 'tt', we now have the AST-equivalents of Python's `repr`, which we denote as $\ulcorner \_ \urcorner^{\mathsf{T}}$ for the type-level add-quote function, and $\ulcorner \_ \urcorner^{\mathsf{t}}$ for the term-level add-quote function. We add constructors `quine`→ and `quine`← that exhibit the isomorphism that defines our type-level quine constructor, though we elide a constructor declaring that these are inverses, as we find it unnecessary.

To construct the proof of Löb's theorem, we need a few other standard constructors, such as '`VAR`$_0$', which references a term in the context; `_‘’ₐ_`, which we use to denote function application; `_‘∘’_`, a function composition operator; and '$\ulcorner$'`VAR`$_0$'$\urcorner^{\mathsf{t}}$', the variant of '`VAR`$_0$' which adds an extra level of syntax-trees. We also include a number of constructors that handle weakening and substitution; this allows us to avoid both inductive-recursive definitions of weakening and substitution, and defining a judgmental equality or conversion relation.

```
  data Term : {Γ : Context} → Type Γ → Set where
    ‘λ’ : ∀ {Γ A B}
      → Term {Γ ▷ A} (W B) → Term (A ‘→’ B)
    ‘tt’ : ∀ {Γ}
      → Term {Γ} ‘⊤’
    ⌜_⌝ᵀ : ∀ {Γ}
      → Type ε
      → Term {Γ} ‘Typeε’
    ⌜_⌝ᵗ : ∀ {Γ T}
```

```
    → Term {ε} T
    → Term {Γ} (‘□’ ‘’ ⌜ T ⌝ᵀ)
  quine→ : ∀ {φ}
    → Term {ε} (Quine φ ‘→’ φ ‘’ ⌜ Quine φ ⌝ᵀ)
  quine← : ∀ {φ}
    → Term {ε} (φ ‘’ ⌜ Quine φ ⌝ᵀ ‘→’ Quine φ)
  -- The constructors below here are for
  -- variables, weakening, and substitution
  ‘VAR₀’ : ∀ {Γ T}
    → Term {Γ ▷ T} (W T)
  _‘’ₐ_ : ∀ {Γ A B}
    → Term {Γ} (A ‘→’ B)
    → Term {Γ} A
    → Term {Γ} B
  _‘∘’_ : ∀ {Γ A B C}
    → Term {Γ} (B ‘→’ C)
    → Term {Γ} (A ‘→’ B)
    → Term {Γ} (A ‘→’ C)
  ‘⌜‘VAR₀’⌝ᵗ’ : ∀ {T}
    → Term {ε ▷ ‘□’ ‘’ ⌜ T ⌝ᵀ}
      (W (‘□’ ‘’ ⌜ ‘□’ ‘’ ⌜ T ⌝ᵀ ⌝ᵀ))
  →SW₁SV→W : ∀ {Γ T X A B} {x : Term {Γ} X}
    → Term (T ‘→’ (W₁ A ‘’ ‘VAR₀’ ‘→’ W B) ‘’ x)
    → Term (T ‘→’ A ‘’ x ‘→’ B)
  ←SW₁SV→W : ∀ {Γ T X A B} {x : Term {Γ} X}
    → Term ((W₁ A ‘’ ‘VAR₀’ ‘→’ W B) ‘’ x ‘→’ T)
    → Term ((A ‘’ x ‘→’ B) ‘→’ T)
  w : ∀ {Γ A T} → Term A → Term {Γ ▷ T} (W A)
  w→ : ∀ {Γ A B X}
    → Term {Γ} (A ‘→’ B)
    → Term {Γ ▷ X} (W A ‘→’ W B)
  _w‘’’’ₐ_ : ∀ {A B T}
    → Term {ε ▷ T} (W (‘□’ ‘’ ⌜ A ‘→’ B ⌝ᵀ))
    → Term {ε ▷ T} (W (‘□’ ‘’ ⌜ A ⌝ᵀ))
    → Term {ε ▷ T} (W (‘□’ ‘’ ⌜ B ⌝ᵀ))

□ : Type ε → Set _
□ = Term {ε}
```

To verify the soundness of our syntax, we provide a model for it and an interpretation into that model. We call particular attention to the interpretation of ‘□’, which is just `Term {ε}`; to `Quine φ`, which is the interpretation of $\phi$ applied to `Quine φ`; and to the interpretations of the quine isomorphism functions, which are just the identity functions.

```
max-level : Level
max-level = lzero -- also works for higher levels

mutual
  ⟦_⟧ᶜ : (Γ : Context) → Set (lsuc max-level)
  ⟦ ε ⟧ᶜ = ⊤
  ⟦ Γ ▷ T ⟧ᶜ = Σ ⟦ Γ ⟧ᶜ ⟦ T ⟧ᵀ

  ⟦_⟧ᵀ : ∀ {Γ}
    → Type Γ → ⟦ Γ ⟧ᶜ → Set max-level
  ⟦ ‘Type’ε ⟧ᵀ Γ⇓ = Lifted (Type ε)
  ⟦ ‘□’ ⟧ᵀ Γ⇓ = Lifted (Term {ε} (lower (Σ.snd Γ⇓)))
  ⟦ Quine φ ⟧ᵀ Γ⇓ = ⟦ φ ⟧ᵀ (Γ⇓ , lift (Quine φ))
  -- The rest of the type-level interpretations
  -- are the obvious ones, if a bit obscured by
  -- carrying around the context.
  ⟦ A ‘→’ B ⟧ᵀ Γ⇓ = ⟦ A ⟧ᵀ Γ⇓ → ⟦ B ⟧ᵀ Γ⇓
  ⟦ ‘⊤’ ⟧ᵀ Γ⇓ = ⊤
  ⟦ ‘⊥’ ⟧ᵀ Γ⇓ = ⊥
  ⟦ W T ⟧ᵀ Γ⇓ = ⟦ T ⟧ᵀ (Σ.fst Γ⇓)
  ⟦ W₁ T ⟧ᵀ Γ⇓ = ⟦ T ⟧ᵀ (Σ.fst (Σ.fst Γ⇓) , Σ.snd Γ⇓)
  ⟦ T ‘’ x ⟧ᵀ Γ⇓ = ⟦ T ⟧ᵀ (Γ⇓ , ⟦ x ⟧ᵗ Γ⇓)

  ⟦_⟧ᵗ : ∀ {Γ T}
    → Term {Γ} T → (Γ⇓ : ⟦ Γ ⟧ᶜ) → ⟦ T ⟧ᵀ Γ⇓
  ⟦ ⌜ x ⌝ᵀ ⟧ᵗ Γ⇓ = lift x
  ⟦ ⌜ x ⌝ᵗ ⟧ᵗ Γ⇓ = lift x
  ⟦ quine→ ⟧ᵗ Γ⇓ x = x
  ⟦ quine← ⟧ᵗ Γ⇓ x = x
  -- The rest of the term-level interpretations
  -- are the obvious ones, if a bit obscured by
  -- carrying around the context.
  ⟦ ‘λ’ f ⟧ᵗ Γ⇓ x = ⟦ f ⟧ᵗ (Γ⇓ , x)
  ⟦ ‘tt’ ⟧ᵗ Γ⇓ = tt
  ⟦ ‘VAR₀’ ⟧ᵗ Γ⇓ = Σ.snd Γ⇓
  ⟦ ‘⌜‘VAR₀’⌝ᵗ’ ⟧ᵗ Γ⇓ = lift ⌜ lower (Σ.snd Γ⇓) ⌝ᵗ
  ⟦ g ‘∘’ f ⟧ᵗ Γ⇓ x = ⟦ g ⟧ᵗ Γ⇓ (⟦ f ⟧ᵗ Γ⇓ x)
  ⟦ f ‘’ₐ x ⟧ᵗ Γ⇓ = ⟦ f ⟧ᵗ Γ⇓ (⟦ x ⟧ᵗ Γ⇓)
  ⟦ ←SW₁SV→W f ⟧ᵗ = ⟦ f ⟧ᵗ
  ⟦ →SW₁SV→W f ⟧ᵗ = ⟦ f ⟧ᵗ
  ⟦ w x ⟧ᵗ Γ⇓ = ⟦ x ⟧ᵗ (Σ.fst Γ⇓)
  ⟦ w→ f ⟧ᵗ Γ⇓ = ⟦ f ⟧ᵗ (Σ.fst Γ⇓)
  ⟦ f w‘’’’ₐ x ⟧ᵗ Γ⇓
    = lift (lower (⟦ f ⟧ᵗ Γ⇓) ‘’ₐ lower (⟦ x ⟧ᵗ Γ⇓))
```

To prove Löb's theorem, we must create the sentence "if this sentence is provable, then $X$", and then provide and inhabitant of that type. We can define this sentence, which we call `‘H’`, as the type-level quine at the function $\lambda v.\ \Box v \to$ ‘$X$’. We can then convert back and forth between the types □ `‘H’` and □ `‘H’` → `‘X’` with our quine isomorphism functions, and a bit of quotation magic and function application gives us a term of type □ `‘H’` → □ `‘X’`; this corresponds to the inference of the provability of Santa Claus' existence from the assumption that the sentence is provable. We compose this with the assumption of Löb's theorem, that □ `‘X’` → `‘X’`, to get a term of type □ `‘H’` → `‘X’`, i.e., a term of type `‘H’`; this is the inference that when provability implies truth, Santa Claus exists, and hence that the sentence is provable. Finally, we apply this to its own quotation, obtaining a term of type □ `‘X’`, i.e., a proof that Santa Claus exists.

```
module inner (‘X’ : Type ε)
       (‘f’ : Term {ε} (‘□’ ‘’ ⌜ ‘X’ ⌝ᵀ ‘→’ ‘X’))
  where
  ‘H’ : Type ε
  ‘H’ = Quine (W₁ ‘□’ ‘’ ‘VAR₀’ ‘→’ W ‘X’)

  ‘toH’ : □ ((‘□’ ‘’ ⌜ ‘H’ ⌝ᵀ ‘→’ ‘X’) ‘→’ ‘H’)
  ‘toH’ = ←SW₁SV→W quine←

  ‘fromH’ : □ (‘H’ ‘→’ (‘□’ ‘’ ⌜ ‘H’ ⌝ᵀ ‘→’ ‘X’))
  ‘fromH’ = →SW₁SV→W quine→

  ‘□‘H’→□‘X’’ : □ (‘□’ ‘’ ⌜ ‘H’ ⌝ᵀ ‘→’ ‘□’ ‘’ ⌜ ‘X’ ⌝ᵀ)
  ‘□‘H’→□‘X’’
    = ‘λ’ (w ⌜ ‘fromH’ ⌝ᵗ
      w‘’’’ₐ ‘VAR₀’
      w‘’’’ₐ ‘⌜‘VAR₀’⌝ᵗ’)
```

| $B$ Says \ $A$ Says | Cooperate | Defect |
|---|---|---|
| Cooperate | (1 year, 1 year) | (0 years, 3 years) |
| Defect | (3 years, 0 years) | (2 years, 2 years) |

**Table 2.** The payoff matrix for the prisoner's dilemma; each cell contains (the years $A$ spends in prison, the years $B$ spends in prison).

```
'h' : Term 'H'
'h' = 'toH' ''ₐ ('f' '∘' '□'H'→□'X'')

Löb : □ 'X'
Löb = 'fromH' ''ₐ 'h' ''ₐ ⌜ 'h' ⌝ᵗ

Löb : ∀ {X} → □ ('□' '' ⌜ X ⌝ᵀ '→' X) → □ X
Löb {X} f = inner.Löb X f

⟦_⟧ : Type ε → Set _
⟦ T ⟧ = ⟦ T ⟧ᵀ tt

'¬'_ : ∀ {Γ} → Type Γ → Type Γ
'¬' T = T '→' '⊥'

löb : ∀ {'X'} → □ ('□' '' ⌜ 'X' ⌝ᵀ '→' 'X') → ⟦ 'X' ⟧
löb f = ⟦_⟧ᵗ (Löb f) tt
```

As above, we can again prove soundness, incompleteness, and non-emptiness.

```
incompleteness : ¬ □ ('¬' ('□' '' ⌜ '⊥' ⌝ᵀ))
incompleteness = löb

soundness : ¬ □ '⊥'
soundness x = ⟦ x ⟧ᵗ tt

non-emptiness : Σ (Type ε) (λ T → □ T)
non-emptiness = '⊤' , 'tt'
```

# 9. Digression: Application of Quining to The Prisoner's Dilemma

In this section, we use a slightly more enriched encoding of syntax; see Appendix B for details.

## 9.1 The Prisoner's Dilemma

The Prisoner's Dilemma is a classic problem in game theory. Two people have been arrested as suspects in a crime and are being held in solatary confinement, with no means of communication. The investigators offer each of them a plea bargain: a decreased sentence for ratting out the other person. Each suspect can then choose to either cooperate with the other suspect by remaining silent, or defect by ratting out the other suspect. The possible outcomes are summarized in Table 2.

Suspect $A$ might reason thusly: "Suppose the other suspect cooperates with me. Then I'd get off with no prison time if I defected, while I'd have to spend a year in prison if I cooperate. Similarly, if the other suspect defects, then I'd get two years in prison for defecting, and three for cooperating. In all cases, I do better by defecting." If suspect $B$ reasons similarly, then both decide to defect, and both get two years in prison, despite the fact that both prefer the (Cooperate, Cooperate) outcome over the (Defect, Defect) outcome!

## 9.2 Adding Source Code

We have the intuition that if both suspects are good at reasoning, and both know that they'll reason the same way, then they should be able to mutually cooperate. One way to formalize this is to talk about programs (rather than people) playing the prisoner's dilemma, and to allow each program access to its own source code and its opponent's source code (Barasz et al. 2014).

We have formalized this framework in Agda: we use `'Bot'` to denote the type of programs that can play in such a prisoner's dilemma; each one takes in source code for two `'Bot'`s and outputs a proposition which is true (a type which is inhabited) if and only if it cooperates with its opponent. Said another way, the output of each bot is a proposition describing the assertion that it cooperates with its opponent.

```
open lob

-- 'Bot' is defined as the fixed point of
-- 'Bot'
-- ↔ (Term 'Bot' → Term 'Bot' → 'Type')
'Bot' : ∀ {Γ} → Type Γ
'Bot' {Γ}
  = Quine (W₁ 'Term' '' 'VAR₀'
      '→' W₁ 'Term' '' 'VAR₀'
      '→' W ('Type' Γ))
```

To construct an executable bot, we could do a bounded search for proofs of this proposition; one useful method described in (Barasz et al. 2014) is to use Kripke frames. This computation is, however, beyond the scope of this paper.

The assertion that a bot $\mathsf{b}_1$ cooperates with a bot $\mathsf{b}_2$ is the result of interpreting the source code for the bot, and feeding the resulting function the source code for $\mathsf{b}_1$ and $\mathsf{b}_2$.

```
-- N.B. "□" means "Term {ε}", i.e., a term in
-- the empty context
_cooperates-with_ : □ 'Bot' → □ 'Bot' → Type ε
b₁ cooperates-with b₂ = lower (⟦ b₁ ⟧ᵗ tt (lift b₁) (lift b₂))
```

We now provide a convenience constructor for building bots, based on the definition of quines, and present four relatively simple bots: DefectBot, CooperateBot, FairBot, and PrudentBot.

```
make-bot : ∀ {Γ}
  → Term {Γ ▷ '□' 'Bot' ▷ W ('□' 'Bot')}
      (W (W ('Type' Γ)))
  → Term {Γ} 'Bot'
make-bot t
  = ←SW₁SV→SW₁SV→W
      quine← ''ₐ 'λ' (→w ('λ' t))

'DefectBot'    : □ 'Bot'
'CooperateBot' : □ 'Bot'
'FairBot'      : □ 'Bot'
'PrudentBot'   : □ 'Bot'
```

The first two bots are very simple: DefectBot never cooperates (the assertion that DefectBot cooperates is a contradiction), while CooperateBot always cooperates. We define these bots, and prove that DefectBot never cooperates and CooperateBot always cooperates.

```
'DefectBot'  = make-bot (w (w ⌜ '⊥' ⌝ᵀ))
'CooperateBot' = make-bot (w (w ⌜ '⊤' ⌝ᵀ))

DB-defects : ∀ {b}
  → ¬ ⟦ 'DefectBot' cooperates-with b ⟧
DB-defects {b} pf = pf

CB-cooperates : ∀ {b}
```

```
    → ⟦ 'CooperateBot' cooperates-with b ⟧
  CB-cooperates {b} = tt
```

We can do better than DefectBot, though, now that we have source code. FairBot cooperates with you if and only if it can find a proof that you cooperate with FairBot. By Löb's theorem, to prove that FairBot cooperates with itself, it sufficies to prove that if there is a proof that FairBot cooperates with itself, then FairBot does, in fact, cooperate with itself. This is obvious, though: FairBot decides whether or not to cooperate with itself by searching for a proof that it does, in fact, cooperate with itself.

To define FairBot, we first define what it means for the other bot to cooperate with some particular bot.

```
-- We can "evaluate" a bot to turn it into a
-- function accepting the source code of two
-- bots.
'eval-bot' : ∀ {Γ}
    → Term {Γ} ('Bot'
      '→' ('□' 'Bot' '→' '□' 'Bot' '→' 'Type' Γ))
'eval-bot' = →SW₁SV→SW₁SV→W quine→

-- We can quote this, and get a function that
-- takes the source code for a bot, and
-- outputs the source code for a function that
-- takes (the source code for) that bot's
-- opponent, and returns an assertion of
-- cooperation with that opponent
``eval-bot'' : ∀ {Γ}
    → Term {Γ} ('□' 'Bot'
      '→' '□' ({- other -} '□' 'Bot' '→' 'Type' Γ))
``eval-bot''
    = 'λ' (w ⌜ 'eval-bot' ⌝ᵗ
      w'''ₐ 'VAR₀'
      w'''ₐ '⌜'VAR₀'⌝ᵗ')

-- The assertion "our opponent cooperates with
-- a bot b" is equivalent to the evalution of
-- our opponent, applied to b. Most of the
-- noise in this statement is manipulation of
-- weakening and substiution.
'other-cooperates-with' : ∀ {Γ}
    → Term {Γ
      ▷ '□' 'Bot'
      ▷ W ('□' 'Bot')}
      (W (W ('□' 'Bot')) '→' W (W ('□' ('Type' Γ))))
'other-cooperates-with' {Γ}
    = 'eval-other''
      '∘' w→ (w (w→ (w ('λ' '⌜'VAR₀'⌝ᵗ'))))
  where
    'eval-other'
      : Term {Γ ▷ '□' 'Bot' ▷ W ('□' 'Bot')}
        (W (W ('□' ('□' 'Bot' '→' 'Type' Γ))))
    'eval-other'
      = w→ (w (w→ (w ``eval-bot''))) ''ₐ 'VAR₀'

    'eval-other''
      : Term (W (W ('□' ('□' 'Bot'))))
        '→' W (W ('□' ('Type' Γ))))
    'eval-other''
      = ww→ (w→ (w (w→ (w '''ₐ'))) ''ₐ 'eval-other')

-- A bot gets its own source code as the first
-- argument (of two)
'self' : ∀ {Γ}
    → Term {Γ ▷ '□' 'Bot' ▷ W ('□' 'Bot')}
      (W (W ('□' 'Bot')))
'self' = w 'VAR₀'

-- A bot gets its opponent's source code as
-- the second argument (of two)
'other' : ∀ {Γ}
    → Term {Γ ▷ '□' 'Bot' ▷ W ('□' 'Bot')}
      (W (W ('□' 'Bot')))
'other' = 'VAR₀'

-- FairBot is the bot that cooperates iff its
-- opponent cooperates with it
'FairBot'
    = make-bot (``□'' ('other-cooperates-with' ''ₐ 'self'))
```

We leave the proof that this formalization of FairBot cooperates with itself as an exercise for the reader. In Appendix C, we present an alternative formalization with a simple proof that FairBot cooperates with itself, but with no general definition of the type of bots; we relegate this code to an appendix so as to not confuse the reader by introducing a different way of handling contexts and weakening in the middle of this paper.

We now come to the final bot: PrudentBot. You do better in the prisoner's dilemma if you cooperate whenever that's required for mutual cooperation, and you defect whenever your opponent would cooperate even if you defected. PrudentBot formalizes an approximation to this intuition: PrudentBot cooperates with you if and only if it can prove that you cooperate with it, and it can prove that you defect against DefectBot (when it assumes that DefectBot does not cooperate with you).

PrudentBot defects against DefectBot. Since there is no proof of $\bot$, PrudentBot does not find a proof that DefectBot cooperates with it, and so it will not cooperate with DefectBot.

By Löb's theorem, PrudentBot cooperates with itself. Under the assumption that $\bot$ is unprovable, PrudentBot can prove that it defects against DefectBot. If we further assume that PrudentBot can find a proof that it cooperates with itself (which we are allowed to do when proving the hypothesis of Löb's theorem), then PrudentBot will, in fact, cooperate with itself. Hence, by Löb's theorem, we can prove that PrudentBot will cooperate with itself.

We leave the formalization of this proof to the reader, and present only the definition of PrudentBot.

```
  -- Convenience notation for triply quoted
  -- negation in a context with at least two
  -- terms
  '''¬''' _ : ∀ {Γ A B}
      → Term {Γ ▷ A ▷ B} (W (W ('□' ('Type' Γ))))
      → Term {Γ ▷ A ▷ B} (W (W ('□' ('Type' Γ))))
  '''¬''' T = T '''→''' w (w ⌜ ⌜ '⊥' ⌝ᵀ ⌝ᵗ)

  -- PrudentBot cooperates if its opponent
  -- cooperates with PrudentBot, and if, under
  -- the assumption that ⊥ is unprovable (¬□⊥),
  -- its opponent does not cooperate with
  -- DefectBot
  'PrudentBot'
      = make-bot (``□''
        (('other-cooperates-with' ''ₐ 'self')
          '''×'''
          (¬□⊥ '''→''' other-defects-against-DB)))
    where
      other-defects-against-DB
```

```
    : Term {_ ▷ '□' 'Bot' ▷ W ('□' 'Bot')}
      (W (W ('□' ('Type' _))))
other-defects-against-DB
  = "'¬"'
      ('other-cooperates-with'
      ''ₐ w (w ⌜ 'DefectBot' ⌝ᵗ))

¬□⊥ : ∀ {Γ A B}
  → Term {Γ ▷ A ▷ B} (W (W ('□' ('Type' Γ))))
¬□⊥ = w (w ⌜ ⌜ '¬' ('□' '⊥') ⌝ᵀ ⌝ᵗ)
```

## 10. Encoding with Add-Quote Function

Now we return to our proving of Löb's theorem. Included in the artifact for this paper[3] is code that replaces the `Quine` constructor with simpler constructors. Because the lack of $\beta$-reduction in the syntax clouds the main points and makes the code rather verbose, we do not include the code in the paper, and instead describe the most interesting and central points.

Recall our Python quine from section 2:

```
(lambda T: □ (T % repr(T)) → X)
 ("(lambda T: □ (T %% repr(T)) → X)\n (%s)")
```

To tranlate this into Agda, we need to give a type to `T`. Clearly, `T` needs to be of type `Type ???` for some context `???`. Since we need to be able to subsitute something into that context, we must have `T : Type (Γ ▷ ???)`, i.e., `T` must be a syntax tree for a type, with a hole in it.

What's the shape of the thing being subtituted? Well, it's a syntax tree for a type with a hole in it… Uh-oh. Our quine's type, naïvely, is infinite!

We know of two ways to work around this. Classical mathematics, which uses Gödel codes instead of abstract syntax trees, uses an untyped representation of proofs. It's only later in the proof of Löb's theorem that a notion of a formula being "well-formed" is introduced.

Here, we describe an alternate approach. Rather than giving up types all-together, we can "box" the type of the hole, to hide it. Using `fst` and `snd` to denote projections from a $\Sigma$ type, using `⌜ A ⌝` to denote the abstract syntax tree for `A`,[4] and using `%s` to denote the first variable in the context (written as `'VAR`$_0$`'` in previous formalizations above), we can write:

```
dummy : Type (ε ▷ ⌜Σ Context Type⌝)
repr : Σ Context Type → Term {ε} ⌜Σ Context Type⌝

cast-fst
  : Σ Context Type → Type (ε ▷ ⌜Σ Context Type⌝)
cast-fst (ε ▷ ⌜Σ Context Type⌝ , T) = T
cast-fst (_ , _) = dummy

LöbSentence : Type ε
LöbSentence
  = (λ (T : Σ Context Type)
       → □ (cast-fst T % repr T) '→' X)
     ( ε ▷ ⌜Σ Context Type⌝
     , ⌜ (λ (T : Σ Context Type)
            → □ (cast-fst T % repr T) '→' X)
          (%s) ⌝
```

[3] In `lob-build-quine.lagda`.

[4] Note that ⌜_⌝ would not be a function in the language, but a meta-level operation.

In this pseudo-Agda code, `cast-fst` unboxes the sentence that it gets, and returns it if it is the right type. Since the sentence is, in fact, always the right type, what we do in the other cases doesn't matter.

Summing up, the key ingredients to this construction are:

- A type of syntactic terms indexed over a type of syntactic types (and contexts)
- Decidable equality on syntactic contexts at a particular point ($\Sigma$ `Context Type`), with appropriate reduction on equal things
- $\Sigma$ types, projections, and appropriate reduction on their projections
- Function types
- A function `repr` which adds a level of quotation to any syntax tree
- Syntax trees for all of the above

In any formalization of dependent type theory with all of these ingredients, we can prove Löb's theorem.

## 11. Conclusion

What remains to be done is formalizing Martin-Löf type theory without assuming `repr` nor assuming a constructor for the type of syntax trees (`'Context'`, `'Type'`, and `'Term'` or `'□'` in our formalizations), and instead support inductive types, and construct these things inductively, and by folding over the inductive definitions there-in.

If you take away only three things from this paper, take away these:

1. There will always be some true things which are not possible to say, no matter how good you are at talking in type theory about type theory.
2. Giving meaning to syntax in a way that doesn't use cases inside cases allows you to talk about when it's okay to add new syntax.
3. If believing in something is enough to make it true, then it already is. Dream big.

## A. Standard Data-Type Declarations

```
open import Agda.Primitive public
  using (Level; _⊔_; lzero; lsuc)

infixl 1 _,_
infixr 2 _×_
infixl 1 _≡_

record ⊤ {ℓ} : Set ℓ where
  constructor tt

data ⊥ {ℓ} : Set ℓ where

¬_ : ∀ {ℓ ℓ'} → Set ℓ → Set (ℓ ⊔ ℓ')
¬_ {ℓ} {ℓ'} T = T → ⊥ {ℓ'}

record Σ {a p} (A : Set a) (P : A → Set p)
       : Set (a ⊔ p)
  where
  constructor _,_
  field
    fst : A
    snd : P fst
```

```
open Σ public

data _+_ {i j} (A : Set i) (B : Set j) : Set (i ⊔ j) where
  inl : A → A + B
  inr : B → A + B

if+ : ∀{A B C : Set} → (A → C) → (B → C) → (A + B → C)
if+ l r (inl x) = l x
if+ l r (inr x) = r x

data Lifted {a b} (A : Set a) : Set (b ⊔ a) where
  lift : A → Lifted A

lower : ∀ {a b A} → Lifted {a} {b} A → A
lower (lift x) = x

_×_ : ∀ {ℓ ℓ'} (A : Set ℓ) (B : Set ℓ') → Set (ℓ ⊔ ℓ')
A × B = Σ A (λ _ → B)

data _≡_ {ℓ} {A : Set ℓ} (x : A) : A → Set ℓ where
  refl : x ≡ x

sym : {A : Set} → {x : A} → {y : A} → x ≡ y → y ≡ x
sym refl = refl

trans : {A : Set} → {x y z : A} → x ≡ y → y ≡ z → x ≡ z
trans refl refl = refl

transport : ∀ {A : Set} {x : A} {y : A} → (P : A → Set)
  → x ≡ y → P x → P y
transport P refl v = v

data List (A : Set) : Set where
  ε : List A
  _::_ : A → List A → List A

_++_ : ∀{A} → List A → List A → List A
ε ++ ys = ys
(x :: xs) ++ ys = x :: (xs ++ ys)
```

## B. Encoding of Löb's Theorem for the Prisoner's Dilemma

```
module lob where
  infixl 2 _▷_
  infixl 3 _‘’_
  infixr 1 _‘→’_
  infixr 1 _‘‘→’’_
  infixr 1 _‘‘‘→’’’_
  infixl 3 _‘’ₐ_
  infixl 3 _w‘‘’’ₐ_
  infixr 2 _‘∘’_
  infixr 2 _‘×’_
  infixr 2 _‘‘×’’_
  infixr 2 _w‘‘×’’_

  mutual
    data Context : Set where
      ε : Context
      _▷_ : (Γ : Context) → Type Γ → Context

    data Type : Context → Set where
      ‘⊤’ : ∀ {Γ} → Type Γ
      ‘⊥’ : ∀ {Γ} → Type Γ
      _‘→’_ : ∀ {Γ} → Type Γ → Type Γ → Type Γ
      _‘×’_ : ∀ {Γ} → Type Γ → Type Γ → Type Γ
      ‘Type’ : ∀ Γ → Type Γ
      ‘Term’ : ∀ {Γ} → Type (Γ ▷ ‘Type’ Γ)
      Quine : ∀ {Γ} → Type (Γ ▷ ‘Type’ Γ) → Type Γ
      W : ∀ {Γ A} → Type Γ → Type (Γ ▷ A)
      W₁ : ∀ {Γ A B}
        → Type (Γ ▷ B)
        → Type (Γ ▷ A ▷ W B)
      _‘’_ : ∀ {Γ A}
        → Type (Γ ▷ A)
        → Term A
        → Type Γ

    data Term : {Γ : Context} → Type Γ → Set where
      ‘tt’ : ∀ {Γ} → Term {Γ} ‘⊤’
      ‘λ’ : ∀ {Γ A B}
        → Term {Γ ▷ A} (W B)
        → Term (A ‘→’ B)
      ‘VAR₀’ : ∀ {Γ T} → Term {Γ ▷ T} (W T)
      ⌜_⌝ᵀ : ∀ {Γ}
        → Type Γ
        → Term {Γ} (‘Type’ Γ)
      ⌜_⌝ᵗ : ∀ {Γ T}
        → Term {Γ} T
        → Term {Γ} (‘Term’ ‘’ ⌜ T ⌝ᵀ)
      ‘⌜‘VAR₀’⌝ᵗ’ : ∀ {Γ T}
        → Term {Γ ▷ ‘Term’ ‘’ ⌜ T ⌝ᵀ}
          (W (‘Term’ ‘’ ⌜ ‘Term’ ‘’ ⌜ T ⌝ᵀ ⌝ᵀ))
      ‘⌜‘VAR₀’⌝ᵀ’ : ∀ {Γ}
        → Term {Γ ▷ ‘Type’ Γ}
          (W (‘Term’ ‘’ ⌜ ‘Type’ Γ ⌝ᵀ))
      _‘’ₐ_ : ∀ {Γ A B}
        → Term {Γ} (A ‘→’ B)
        → Term {Γ} A
        → Term {Γ} B
      ‘‘×’’’ : ∀ {Γ}
        → Term {Γ} (‘Type’ Γ
          ‘→’ ‘Type’ Γ
          ‘→’ ‘Type’ Γ)
      quine→ : ∀ {Γ φ}
        → Term {Γ}
          (Quine φ ‘→’ φ ‘’ ⌜ Quine φ ⌝ᵀ)
      quine← : ∀ {Γ φ}
        → Term {Γ}
          (φ ‘’ ⌜ Quine φ ⌝ᵀ ‘→’ Quine φ)
      SW : ∀ {Γ X A} {a : Term A}
        → Term {Γ} (W X ‘’ a)
        → Term X
      →SW₁SV→W
        : ∀ {Γ T X A B} {x : Term X}
        → Term {Γ}
          (T ‘→’ (W₁ A ‘’ ‘VAR₀’ ‘→’ W B) ‘’ x)
        → Term {Γ}
          (T ‘→’ A ‘’ x ‘→’ B)
      ←SW₁SV→W
        : ∀ {Γ T X A B} {x : Term X}
        → Term {Γ}
```

```
      ((W₁ A ‘’ ‘VAR₀’ ‘→’ W B) ‘’ x ‘→’ T)
   → Term {Γ}
      ((A ‘’ x ‘→’ B) ‘→’ T)
→SW₁SV→SW₁SV→W
   : ∀ {Γ T X A B} {x : Term X}
   → Term {Γ} (T ‘→’ (W₁ A ‘’ ‘VAR₀’
      ‘→’ W₁ A ‘’ ‘VAR₀’
      ‘→’ W B) ‘’ x)
   → Term {Γ} (T ‘→’ A ‘’ x ‘→’ A ‘’ x ‘→’ B)
←SW₁SV→SW₁SV→W
   : ∀ {Γ T X A B} {x : Term X}
   → Term {Γ} ((W₁ A ‘’ ‘VAR₀’
      ‘→’ W₁ A ‘’ ‘VAR₀’
      ‘→’ W B) ‘’ x
      ‘→’ T)
   → Term {Γ} ((A ‘’ x ‘→’ A ‘’ x ‘→’ B) ‘→’ T)
w : ∀ {Γ A T}
   → Term {Γ} A
   → Term {Γ ▷ T} (W A)
w→ : ∀ {Γ A B X}
   → Term {Γ ▷ X} (W (A ‘→’ B))
   → Term {Γ ▷ X} (W A ‘→’ W B)
→w : ∀ {Γ A B X}
   → Term {Γ ▷ X} (W A ‘→’ W B)
   → Term {Γ ▷ X} (W (A ‘→’ B))
ww→ : ∀ {Γ A B X Y}
   → Term {Γ ▷ X ▷ Y} (W (W (A ‘→’ B)))
   → Term {Γ ▷ X ▷ Y} (W (W A) ‘→’ W (W B))
→ww : ∀ {Γ A B X Y}
   → Term {Γ ▷ X ▷ Y} (W (W A) ‘→’ W (W B))
   → Term {Γ ▷ X ▷ Y} (W (W (A ‘→’ B)))
_‘∘’_ : ∀ {Γ A B C}
   → Term {Γ} (B ‘→’ C)
   → Term {Γ} (A ‘→’ B)
   → Term {Γ} (A ‘→’ C)
_w‘’’ₐ_ : ∀ {Γ A B T}
   → Term {Γ ▷ T} (W (‘Term’ ‘’ ⌜ A ‘→’ B ⌝ᵀ))
   → Term {Γ ▷ T} (W (‘Term’ ‘’ ⌜ A ⌝ᵀ))
   → Term {Γ ▷ T} (W (‘Term’ ‘’ ⌜ B ⌝ᵀ))
‘’’ₐ’ : ∀ {Γ A B}
   → Term {Γ} (‘Term’ ‘’ ⌜ A ‘→’ B ⌝ᵀ
      ‘→’ ‘Term’ ‘’ ⌜ A ⌝ᵀ
      ‘→’ ‘Term’ ‘’ ⌜ B ⌝ᵀ)
‘’□’’ : ∀ {Γ A B}
   → Term {Γ ▷ A ▷ B}
      (W (W (‘Term’ ‘’ ⌜ ‘Type’ Γ ⌝ᵀ)))
   → Term {Γ ▷ A ▷ B}
      (W (W (‘Type’ Γ)))
_‘’→’’_ : ∀ {Γ}
   → Term {Γ} (‘Type’ Γ)
   → Term {Γ} (‘Type’ Γ)
   → Term {Γ} (‘Type’ Γ)
_‘’‘→’’’_ : ∀ {Γ A B}
   → Term {Γ ▷ A ▷ B}
      (W (W (‘Term’ ‘’ ⌜ ‘Type’ Γ ⌝ᵀ)))
   → Term {Γ ▷ A ▷ B}
      (W (W (‘Term’ ‘’ ⌜ ‘Type’ Γ ⌝ᵀ)))
   → Term {Γ ▷ A ▷ B}
      (W (W (‘Term’ ‘’ ⌜ ‘Type’ Γ ⌝ᵀ)))
_‘’×’’_ : ∀ {Γ A B}
   → Term {Γ ▷ A ▷ B}
      (W (W (‘Term’ ‘’ ⌜ ‘Type’ Γ ⌝ᵀ)))
   → Term {Γ ▷ A ▷ B}
      (W (W (‘Term’ ‘’ ⌜ ‘Type’ Γ ⌝ᵀ)))
   → Term {Γ ▷ A ▷ B}
      (W (W (‘Term’ ‘’ ⌜ ‘Type’ Γ ⌝ᵀ)))

□ : Type ε → Set _
□ = Term {ε}

‘□’ : ∀ {Γ} → Type Γ → Type Γ
‘□’ T = ‘Term’ ‘’ ⌜ T ⌝ᵀ

_‘’×’’_ : ∀ {Γ}
   → Term {Γ} (‘Type’ Γ)
   → Term {Γ} (‘Type’ Γ)
   → Term {Γ} (‘Type’ Γ)
A ‘’×’’ B = ‘’×’’’ ‘’ₐ A ‘’ₐ B

max-level : Level
max-level = lzero

mutual
  ⟦_⟧ᶜ : (Γ : Context) → Set (lsuc max-level)
  ⟦ ε ⟧ᶜ = ⊤
  ⟦ Γ ▷ T ⟧ᶜ = Σ ⟦ Γ ⟧ᶜ ⟦ T ⟧ᵀ

  ⟦_⟧ᵀ : {Γ : Context}
     → Type Γ
     → ⟦ Γ ⟧ᶜ
     → Set max-level
  ⟦ W T ⟧ᵀ ⟦Γ⟧
     = ⟦ T ⟧ᵀ (Σ.fst ⟦Γ⟧)
  ⟦ W₁ T ⟧ᵀ ⟦Γ⟧
     = ⟦ T ⟧ᵀ (Σ.fst (Σ.fst ⟦Γ⟧) , Σ.snd ⟦Γ⟧)
  ⟦ T ‘’ x ⟧ᵀ ⟦Γ⟧ = ⟦ T ⟧ᵀ (⟦Γ⟧ , ⟦ x ⟧ᵗ ⟦Γ⟧)
  ⟦ ‘Type’ Γ ⟧ᵀ ⟦Γ⟧
     = Lifted (Type Γ)
  ⟦ ‘Term’ ⟧ᵀ ⟦Γ⟧
     = Lifted (Term (lower (Σ.snd ⟦Γ⟧)))
  ⟦ A ‘→’ B ⟧ᵀ ⟦Γ⟧ = ⟦ A ⟧ᵀ ⟦Γ⟧ → ⟦ B ⟧ᵀ ⟦Γ⟧
  ⟦ A ‘×’ B ⟧ᵀ ⟦Γ⟧ = ⟦ A ⟧ᵀ ⟦Γ⟧ × ⟦ B ⟧ᵀ ⟦Γ⟧
  ⟦ ‘⊤’ ⟧ᵀ ⟦Γ⟧ = ⊤
  ⟦ ‘⊥’ ⟧ᵀ ⟦Γ⟧ = ⊥
  ⟦ Quine φ ⟧ᵀ ⟦Γ⟧ = ⟦ φ ⟧ᵀ (⟦Γ⟧ , (lift (Quine φ)))

  ⟦_⟧ᵗ : ∀ {Γ : Context} {T : Type Γ}
     → Term T
     → (⟦Γ⟧ : ⟦ Γ ⟧ᶜ)
     → ⟦ T ⟧ᵀ ⟦Γ⟧
  ⟦ ⌜ x ⌝ᵀ ⟧ᵗ ⟦Γ⟧ = lift x
  ⟦ ⌜ x ⌝ᵗ ⟧ᵗ ⟦Γ⟧ = lift x
  ⟦ ‘⌜‘VAR₀’⌝ᵗ’ ⟧ᵗ ⟦Γ⟧
     = lift ⌜ lower (Σ.snd ⟦Γ⟧) ⌝ᵗ
  ⟦ ‘⌜‘VAR₀’⌝ᵀ’ ⟧ᵗ ⟦Γ⟧
     = lift ⌜ lower (Σ.snd ⟦Γ⟧) ⌝ᵀ
  ⟦ f ‘’ₐ x ⟧ᵗ ⟦Γ⟧ = ⟦ f ⟧ᵗ ⟦Γ⟧ (⟦ x ⟧ᵗ ⟦Γ⟧)
  ⟦ ‘tt’ ⟧ᵗ ⟦Γ⟧ = tt
  ⟦ quine→ {φ} ⟧ᵗ ⟦Γ⟧ x = x
  ⟦ quine← {φ} ⟧ᵗ ⟦Γ⟧ x = x
  ⟦ ‘λ’ f ⟧ᵗ ⟦Γ⟧ x = ⟦ f ⟧ᵗ (⟦Γ⟧ , x)
```

```
⟦ 'VAR₀' ⟧ᵗ ⟦Γ⟧ = Σ.snd ⟦Γ⟧
⟦ SW t ⟧ᵗ = ⟦ t ⟧ᵗ
⟦ ←SW₁SV→W f ⟧ᵗ = ⟦ f ⟧ᵗ
⟦ →SW₁SV→W f ⟧ᵗ = ⟦ f ⟧ᵗ
⟦ ←SW₁SV→SW₁SV→W f ⟧ᵗ = ⟦ f ⟧ᵗ
⟦ →SW₁SV→SW₁SV→W f ⟧ᵗ = ⟦ f ⟧ᵗ
⟦ w x ⟧ᵗ ⟦Γ⟧ = ⟦ x ⟧ᵗ (Σ.fst ⟦Γ⟧)
⟦ w→ f ⟧ᵗ ⟦Γ⟧ = ⟦ f ⟧ᵗ ⟦Γ⟧
⟦ →w f ⟧ᵗ ⟦Γ⟧ = ⟦ f ⟧ᵗ ⟦Γ⟧
⟦ ww→ f ⟧ᵗ ⟦Γ⟧ = ⟦ f ⟧ᵗ ⟦Γ⟧
⟦ →ww f ⟧ᵗ ⟦Γ⟧ = ⟦ f ⟧ᵗ ⟦Γ⟧
⟦ "×"' ⟧ᵗ ⟦Γ⟧ A B = lift (lower A '×' lower B)
⟦ g '∘' f ⟧ᵗ ⟦Γ⟧ x = ⟦ g ⟧ᵗ ⟦Γ⟧ (⟦ f ⟧ᵗ ⟦Γ⟧ x)
⟦ f w"'"ₐ x ⟧ᵗ ⟦Γ⟧
  = lift (lower (⟦ f ⟧ᵗ ⟦Γ⟧) "'ₐ lower (⟦ x ⟧ᵗ ⟦Γ⟧))
⟦ "'ₐ' ⟧ᵗ ⟦Γ⟧ f x
  = lift (lower f "'ₐ lower x)
⟦ "□"' {Γ} T ⟧ᵗ ⟦Γ⟧
  = lift ('Term' "' lower (⟦ T ⟧ᵗ ⟦Γ⟧))
⟦ A "→"' B ⟧ᵗ ⟦Γ⟧
  = lift
      (lower (⟦ A ⟧ᵗ ⟦Γ⟧) '→' lower (⟦ B ⟧ᵗ ⟦Γ⟧))
⟦ A "'→"' B ⟧ᵗ ⟦Γ⟧
  = lift
      (lower (⟦ A ⟧ᵗ ⟦Γ⟧) "'→"' lower (⟦ B ⟧ᵗ ⟦Γ⟧))
⟦ A "'×"' B ⟧ᵗ ⟦Γ⟧
  = lift
      (lower (⟦ A ⟧ᵗ ⟦Γ⟧) "'×"' lower (⟦ B ⟧ᵗ ⟦Γ⟧))


module inner (X' : Type ε)
  (f' : Term {ε} ('□' X' '→' X'))
where
  'H' : Type ε
  'H' = Quine (W₁ 'Term' "' 'VAR₀' '→' W X')

  'toH' : □ (('□' 'H' '→' X') '→' 'H')
  'toH' = ←SW₁SV→W quine←

  'fromH' : □ ('H' '→' ('□' 'H' '→' X'))
  'fromH' = →SW₁SV→W quine→

  '□'H'→□'X"' : □ ('□' 'H' '→' '□' X')
  '□'H'→□'X"'
    = 'λ' (w ⌜ 'fromH' ⌝ᵗ
      w"'"ₐ 'VAR₀'
      w"'"ₐ '⌜'VAR₀'⌝ᵗ')

  'h' : Term 'H'
  'h' = 'toH' "'ₐ (f' '∘' '□'H'→□'X"')

  Löb : □ X'
  Löb = 'fromH' "'ₐ 'h' "'ₐ ⌜ 'h' ⌝ᵗ

Löb : ∀ {X}
  → Term {ε} ('□' X '→' X) → Term {ε} X
Löb {X} f = inner.Löb X f

⟦_⟧ : Type ε → Set _
⟦ T ⟧ = ⟦ T ⟧ᵀ tt
```

```
'¬'_ : ∀ {Γ} → Type Γ → Type Γ
'¬' T = T '→' '⊥'

_w"'×"'_ : ∀ {Γ X}
  → Term {Γ ▷ X} (W ('Type' Γ))
  → Term {Γ ▷ X} (W ('Type' Γ))
  → Term {Γ ▷ X} (W ('Type' Γ))
A w"'×"' B = w→ (w→ (w "'×"') "'ₐ A) "'ₐ B

löb : ∀ {'X'} → □ ('□' 'X' '→' 'X') → ⟦ 'X' ⟧
löb f = ⟦ Löb f ⟧ᵗ tt

incompleteness : ¬ □ ('¬' ('□' '⊥'))
incompleteness = löb

soundness : ¬ □ '⊥'
soundness x = ⟦ x ⟧ᵗ tt

non-emptyness : Σ (Type ε) (λ T → □ T)
non-emptyness = '⊤' , 'tt'
```

## C. Self cooperation in STLC

We'll make use of a few particularly useful dependent combinators throughout this section - they're defined below.

```
_∘_ : ∀ {i j k}{A : Set i}{
      → ({x : A} (y : B x) → C y) → (g : (x : A) → B x) (x : A) → C (g x)
f ∘ g = λ x → f (g x)

infixl 8 _ˢ_

_ˢ_ : ∀ {i j k}{A : Set i}{B : A → Set j}{C : (x : A) → B x → Set k}
  → ((x : A) (y : B x) → C x y)
    → (g : (x : A) → B x) (x : A) → C x (g x)
f ˢ g = λ x → f x (g x)

ᵏ : {A B : Set} → A → B → A
ᵏ a b = a

^ : ∀ {i j k}{S : Set i}{T : S → Set j}{P : Σ S T → Set k}
  → ((σ : Σ S T) → P σ)
  → (s : S) (t : T s) → P (s , t)
^ f s t = f (s , t)
```

It turns out that we can define all the things we need to prove self-cooperation of FairBot in a variant of the simply typed lambda calculus. In order to do this, however, we have to define □ somewhat differently. Particularly, we abandon the notion of a unary □ and instead base our theory on a binary operator denoting provability in a context.

```
infixr 5 _⊢_ _'⊢'_
infixr 10 _'→'_ _'+'_ _'×'_

data ⋆ : Set where
  _'⊢'_ : List ⋆ → ⋆ → ⋆
  _'→'_ _'×'_ _'+'_ : ⋆ → ⋆ → ⋆
  '0' '1' : ⋆

Con = List ⋆
```

We will then need some way to handle binding. For reasons of simplicity, we'll make use of a dependent form of DeBrujin variables.

```
data _∈_ (T : ⋆) : Con → Set where
```

First we want our "variable zero", which lets us pick off the "top" element of the context.

```
top : ∀{Γ} → T ∈ (T :: Γ)
```

Then we want a way to extend variables to work in larger contexts.

```
pop : ∀{Γ S} → T ∈ Γ → T ∈ (S :: Γ)
```

And, finally, we are ready to define the term language for our extended STLC.

```
data _⊢_ (Γ : Con) : ⋆ → Set where
```

The next few constructors are fairly standard. Before anything else, we want to be able to lift bindings into terms.

```
var : ∀{T} → T ∈ Γ → Γ ⊢ T
```

Then the intro rules for all of our easier datatypes.

```
<> : Γ ⊢ '1'
_,_ : ∀{A B} → Γ ⊢ A → Γ ⊢ B → Γ ⊢ A '×' B
inl : ∀{A B} → Γ ⊢ A → Γ ⊢ A '+' B
inr : ∀{A B} → Γ ⊢ B → Γ ⊢ A '+' B
'0'-elim : ∀{A} → Γ ⊢ '0' → Γ ⊢ A
'+'-elim : ∀{A B C} → Γ ⊢ (A '→' C) → Γ ⊢ (B '→' C) → Γ ⊢ A '+' B → Γ ⊢ C
π₁ : ∀{A B} → Γ ⊢ A '×' B → Γ ⊢ A
π₂ : ∀{A B} → Γ ⊢ A '×' B → Γ ⊢ B
```

Then, of course, we need to handle function types. Lambdas are handled in essentially the same way as in the dependent formalism, but with simple codomains rather than dependent ones. Similarly, application can avoid the complications of dependent substitution, and just return the concrete codomain type.

```
lam : ∀{A B} → (A :: Γ) ⊢ B → Γ ⊢ (A '→' B)
_#_ : ∀{A B} → Γ ⊢ (A '→' B) → Γ ⊢ A → Γ ⊢ B
```

At this point things become more delicate. To properly capture GL, we want our theory to validate the rules

1. `⊢ A → ⊢ □ A`
2. `⊢ □ A '→' □ □ A`

But *not* `⊢ A '→' □ A`. If we only had the unary □ operator we would run into difficulty later. Crucially, we couldn't add the rule `Γ ⊢ A → Γ ⊢ □ A`, since this would let us prove `A '→' □ A`.

But, luckily enough, we didn't restrict ourselves in this way. Instead we took as primitive the more general notion of provability in a context, and this lets us give a proper account of □ without compromising on strength.

We will denote by Gödel quotes the constructor corresponding to rule 1.

```
⌜_⌝ : ∀{Δ A} → Δ ⊢ A → Γ ⊢ (Δ '⊢' A)
```

Similarly, we will write the rule validating `□ A '→' □ □ A` as `quot`.

```
quot : ∀{Δ A} → Γ ⊢ (Δ '⊢' A) → Γ ⊢ (Δ '⊢' (Δ '⊢' A))
```

We would like to be able to apply functions under □, and for this we introduce the so-called "distribution" rule. In GL it takes the form `⊢ □ (A '→' B) → ⊢ (□ A '→' □ B)`. For us it is not much more complicated.

```
dist : ∀{Δ A B} → Γ ⊢ (Δ '⊢' (A '→' B)) → Γ ⊢ (Δ '⊢' A) → Γ ⊢ (Δ '⊢' B)
```

And, finally, we include the Löbian axiom, translated to the new contextual provability type.

```
Lob : ∀{Δ A} → Γ ⊢ (Δ '⊢' ((Δ '⊢' A) '→' A)) → Γ ⊢ (Δ '⊢' A)
```

We also specify a fixity for function application. `infixl 50 _#_`

From these constructors we can prove the simpler form of the Lob rule.

```
lob : ∀{Γ A} → Γ ⊢ ((Γ '⊢' A) '→' A) → Γ ⊢ A
lob t = t # Lob ⌜ t ⌝
```

Of course, because we are using DeBrujin indices, before we can do too much we'll need to give an account of lifting. Thankfully, unlike when we were dealing with dependent type theory, we can define these computationally, and get for free all the congruences we had to add as axioms before.

Our definition of weakening is unremarkable, but we've included it below as a reference. One surprising and nice aspect of these definitions is that they were fully automatic - each right hand side generated by Agsy without any hinting.

```
lift-var : ∀{Γ A} T Δ → A ∈ (Δ ++ Γ) → A ∈ (Δ ++ (T :: Γ))
lift-var T ε v = pop v
lift-var T (A :: Δ) top = top
lift-var T (x :: Δ) (pop v) = pop (lift-var T Δ v)

lift-tm : ∀{Γ A} T Δ → (Δ ++ Γ) ⊢ A → (Δ ++ (T :: Γ)) ⊢ A
lift-tm T Δ (var x) = var (lift-var T Δ x)
lift-tm T Δ <> = <>
lift-tm T Δ (a , b) = lift-tm T Δ a , lift-tm T Δ b
lift-tm T Δ (inl t) = inl (lift-tm T Δ t)
lift-tm T Δ (inr t) = inr (lift-tm T Δ t)
lift-tm T Δ ('0'-elim t) = '0'-elim (lift-tm T Δ t)
lift-tm T Δ ('+'-elim t t₁ t₂) = '+'-elim (lift-tm T Δ t) (lift-tm T Δ t₁) (lift
lift-tm T Δ (π₁ t) = π₁ (lift-tm T Δ t)
lift-tm T Δ (π₂ t) = π₂ (lift-tm T Δ t)
lift-tm T Δ (lam t) = lam (lift-tm T (_ :: Δ) t)
lift-tm T Δ (t # t₁) = lift-tm T Δ t # lift-tm T Δ t₁
lift-tm T Δ ⌜ t ⌝ = ⌜ t ⌝
lift-tm T Δ (quot t) = quot (lift-tm T Δ t)
lift-tm T Δ (dist t t₁) = dist (lift-tm T Δ t) (lift-tm T Δ t₁)
lift-tm T Δ (Lob t) = Lob (lift-tm T Δ t)
```

Having defined lift-tming at all levels, we can give weakening as a special case.

```
wk : ∀{Γ A B} → Γ ⊢ A → (B :: Γ) ⊢ A
wk = lift-tm _ ε
```

Finally, we define function composition for our internal language.

```
infixl 10 _∘'_
_∘'_ : ∀{Γ A B C} → Γ ⊢ (B '→' C) → Γ ⊢ (A '→' B) → Γ ⊢ (A '→' C)
f ∘' g = lam (wk f # (wk g # var top))
```

Now we are ready to prove that FairBot cooperates with itself. Sadly, our type system isn't expressive enough to give a general type of bots, but we can still prove things about the interactions of particular bots if we substitute their types by hand. For example, we can state the desired theorem (that FairBot cooperates with itself) as:

```
distf : ∀{Γ Δ A B} → Γ ⊢ (Δ '⊢' A '→' B) → Γ ⊢ (Δ '⊢' A) '→' (Δ '⊢' B)
distf bf = lam (dist (wk bf) (var top))

evf : ∀{Γ Δ A} → Γ ⊢ (Δ '⊢' A) '→' (Δ '⊢' (Δ '⊢' A))
evf = lam (quot (var top))

fb-fb-cooperate : ∀{Γ A B} → Γ ⊢ (Γ '⊢' A) '→' B → Γ ⊢(Γ '⊢' B) '→' A
fb-fb-cooperate a b = lob (b ∘' distf ⌜ a ⌝ ∘' evf) , lob (a ∘' distf ⌜ b ⌝ ∘' evf
```

We can also state the theorem in a more familiar form with a couple abbreviations

```
'□' = _'⊢'_ ε
□ = _⊢_ ε

fb-fb-cooperate' : ∀{A B} → □ ('□' A '→' B) → □ ('□' B '→' A) → □ (A
fb-fb-cooperate' = fb-fb-cooperate
```

We'd also like to show all the metatheoretic properites we had before: soundness, inhabitedness, and incompleteness. We can show inhabitedness immediately in several different ways. We'll take the easiest one.

```
inhabited : Σ ⋆ λ T → ε ⊢ T
inhabited = '1' , <>
```

To prove soundness and incompleteness we'll first need to give the standard interpretation. Again, the simplicity of our system

makes our lives easier. We define the interpreter for types as follows:

```
⟦_⟧⋆ : ⋆ → Set
⟦ Δ '⊢' T ⟧⋆ = Δ ⊢ T
⟦ A '→' B ⟧⋆ = ⟦ A ⟧⋆ → ⟦ B ⟧⋆
⟦ A '×' B ⟧⋆ = ⟦ A ⟧⋆ × ⟦ B ⟧⋆
⟦ A '+' B ⟧⋆ = ⟦ A ⟧⋆ + ⟦ B ⟧⋆
⟦ '0' ⟧⋆ = ⊥
⟦ '1' ⟧⋆ = ⊤
```

The interpreter for contexts is simplified - we only need simple products to interpret simple contexts.

```
⟦_⟧c : Con → Set
⟦ ε ⟧c = ⊤
⟦ x :: Γ ⟧c = ⟦ Γ ⟧c × ⟦ x ⟧⋆
```

We can then interpret variables in any interpretable context.

```
⟦_⟧v : ∀{Γ A} → A ∈ Γ → ⟦ Γ ⟧c → ⟦ A ⟧⋆
⟦ top ⟧v = snd
⟦ pop v ⟧v = ⟦ v ⟧v ∘ fst
```

And now we can interpret terms.

```
⟦_⟧t : ∀{Γ A} → Γ ⊢ A → ⟦ Γ ⟧c → ⟦ A ⟧⋆
⟦ var v ⟧t = ⟦ v ⟧v
⟦ <> ⟧t = ᵏ _
⟦ a , b ⟧t = ᵏ _,_ ˢ ⟦ a ⟧t ˢ ⟦ b ⟧t
⟦ inl a ⟧t = ᵏ inl ˢ ⟦ a ⟧t
⟦ inr b ⟧t = ᵏ inr ˢ ⟦ b ⟧t
⟦ '0'-elim t ⟧t = ᵏ (λ ()) ˢ ⟦ t ⟧t
⟦ '+'-elim l r s ⟧t = ᵏ if+ ˢ ⟦ l ⟧t ˢ ⟦ r ⟧t ˢ ⟦ s ⟧t
⟦ π₁ t ⟧t = ᵏ fst ˢ ⟦ t ⟧t
⟦ π₂ t ⟧t = ᵏ snd ˢ ⟦ t ⟧t
⟦ lam b ⟧t = ^ ⟦ b ⟧t
⟦ f # x ⟧t = ⟦ f ⟧t ˢ ⟦ x ⟧t
⟦ ⌜ t ⌝ ⟧t = ᵏ t
⟦ quot t ⟧t = ᵏ ⌜ _ ⌝ ˢ ⟦ t ⟧t
⟦ dist f x ⟧t = ᵏ _#_ ˢ ⟦ f ⟧t ˢ ⟦ x ⟧t
⟦ Lob l ⟧t = ᵏ lob ˢ ⟦ l ⟧t
```

Which lets us prove all our sanity checks.

```
'¬'_ : ⋆ → ⋆
'¬' T = T '→' '0'

consistency : ¬ (□ '0')
consistency f = ⟦ f ⟧t tt

incompleteness : ¬ (□ ('¬' '□' '0'))
incompleteness t = ⟦ lob t ⟧t tt

soundness : ∀{A} → □ A → ⟦ A ⟧⋆
soundness a = ⟦ a ⟧t tt
```

## Acknowledgments

[Redacted for the anonymous submission]

## References


A. W. Appel, P.-A. Mellies, C. D. Richards, and J. Vouillon. A very modal model of a modern, major, general type system. *ACM SIGPLAN Notices*, 42(1):109–122, 2007. URL `https://www.cs.princeton.edu/~appel/papers/modalmodel.pdf`.

M. Barasz, P. Christiano, B. Fallenstein, M. Herreshoff, P. LaVictoire, and E. Yudkowsky. Robust cooperation in the prisoner's dilemma: Program equilibrium via provability logic. *ArXiv e-prints*, Jan 2014. URL `http://arxiv.org/pdf/1401.5577v1.pdf`.

M. Brown and J. Palsberg. Breaking through the normalization barrier: A self-interpreter for f-omega. In *Proceedings of the 43rd Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages*, pages 5–17. ACM, 2016. doi: 10.1145/2837614.2837623. URL `http://compilers.cs.ucla.edu/popl16/popl16-full.pdf`.

J. Chapman. Type theory should eat itself. *Electronic Notes in Theoretical Computer Science*, 228:21–36, 2009. ISSN 1571-0661. doi: http://dx.doi.org/10.1016/j.entcs.2008.12.114. URL `http://www.sciencedirect.com/science/article/pii/S157106610800577X`. Proceedings of the International Workshop on Logical Frameworks and Metalanguages: Theory and Practice (LFMTP 2008).

N. A. Danielsson. A formalisation of a dependently typed language as an inductive-recursive family. In T. Altenkirch and C. McBride, editors, *Types for Proofs and Programs*, volume 4502 of *Lecture Notes in Computer Science*, chapter A Formalisation of a Dependently Typed Language as an Inductive-Recursive Family, pages 93–109. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-74464-1. doi: 10.1007/978-3-540-74464-1_7. URL `http://dx.doi.org/10.1007/978-3-540-74464-1_7`.

C. McBride. Outrageous but meaningful coincidences: Dependent type-safe syntax and evaluation. In *Proceedings of the 6th ACM SIGPLAN workshop on Generic programming*, pages 1–12. ACM, 2010. URL `https://personal.cis.strath.ac.uk/conor.mcbride/pub/DepRep/DepRep.pdf`.

R. O'Connor. Essential incompleteness of arithmetic verified by coq. *CoRR*, abs/cs/0505034, 2005. URL `http://arxiv.org/abs/cs/0505034`.

D. Piponi. From löb's theorem to spreadsheet evaluation, November 2006. URL `http://blog.sigfpe.com/2006/11/from-l-theorem-to-spreadsheet.html`.

G. K. Pullum. Scooping the loop snooper, October 2000. URL `http://www.lel.ed.ac.uk/~gpullum/loopsnoop.html`.

N. Shankar. *Proof-checking Metamathematics (Theorem-proving)*. PhD thesis, The University of Texas at Austin, 1986. AAI8717580.

N. Shankar. *Metamathematics, Machines and Gödel's Proof*. Cambridge University Press, 1997.

B. Yudkowsky. Lob's theorem cured my social anxiety, February 2014. URL `http://agentyduck.blogspot.com/2014/02/lobs-theorem-cured-my-social-anxiety.html`.