Quine's underdetermination thesis

Eric Johannesson

Stockholm University

The Swedish Congress of Philosophy 2022
Lund
June 12, 2022

On the one hand...

Finding empirically equivalent but logically incompatible rivals is easy:

> *Take some theory formulation and select two of its terms, say 'electron' and 'molecule'. I am supposing that these do not figure essentially in any observation sentences; they are purely theoretical. Now let us transform our theory formulation merely by switching these two terms throughout. The new theory formulation will be logically incompatible with the old: it will affirm things about so-called electrons that the other denies.*

W. V. Quine. On empirically equivalent systems of the world. *Erkenntnis*, 9(3):313–328, 1975.

On the other hand...

> *Yet their only difference, the man in the street would say, is terminological; the one theory formulation uses the technical terms 'molecule' and 'electron' to name what the other formulation calls 'electron' and 'molecule'. The two formulations express, he would say, the same theory.*

More generally, according to Quine, two formulations may be taken to express the same theory if they are "reconcilable by reconstrual of predicates", meaning roughly that there is a way of replacing the predicates of one by formulas of the other to obtain a formulation logically equivalent to the latter.

Quine's underdetermination thesis

The thesis of under-determination, even in my latest tempered version, asserts that our system of the world is bound to have empirically equivalent alternatives that are not reconcilable by reconstrual of predicates however devious. This, for me, is an open question.

The underdetermination property

- A theory T has the **underdetermination property** just in case there is a theory T' such that
  - (i) T and T' are empirically equivalent,
  - (ii) T and T' are jointly inconsistent, and
  - (iii) T and T' are not theoretically equivalent.
- Observe that, when it comes to establishing this property for a given theory, it is sufficient to do so under the strongest notion of empirical equivalence in combination with the weakest notion of theoretical equivalence.

Assumptions

- A theory is a set of sentences of a first-order single-sorted language without function symbols.
- The predicates of the language is partitioned into an empirical and a theoretical part.

Relative to such a partition, various notions of *empirical equivalence* can be defined.

Empirical equivalence

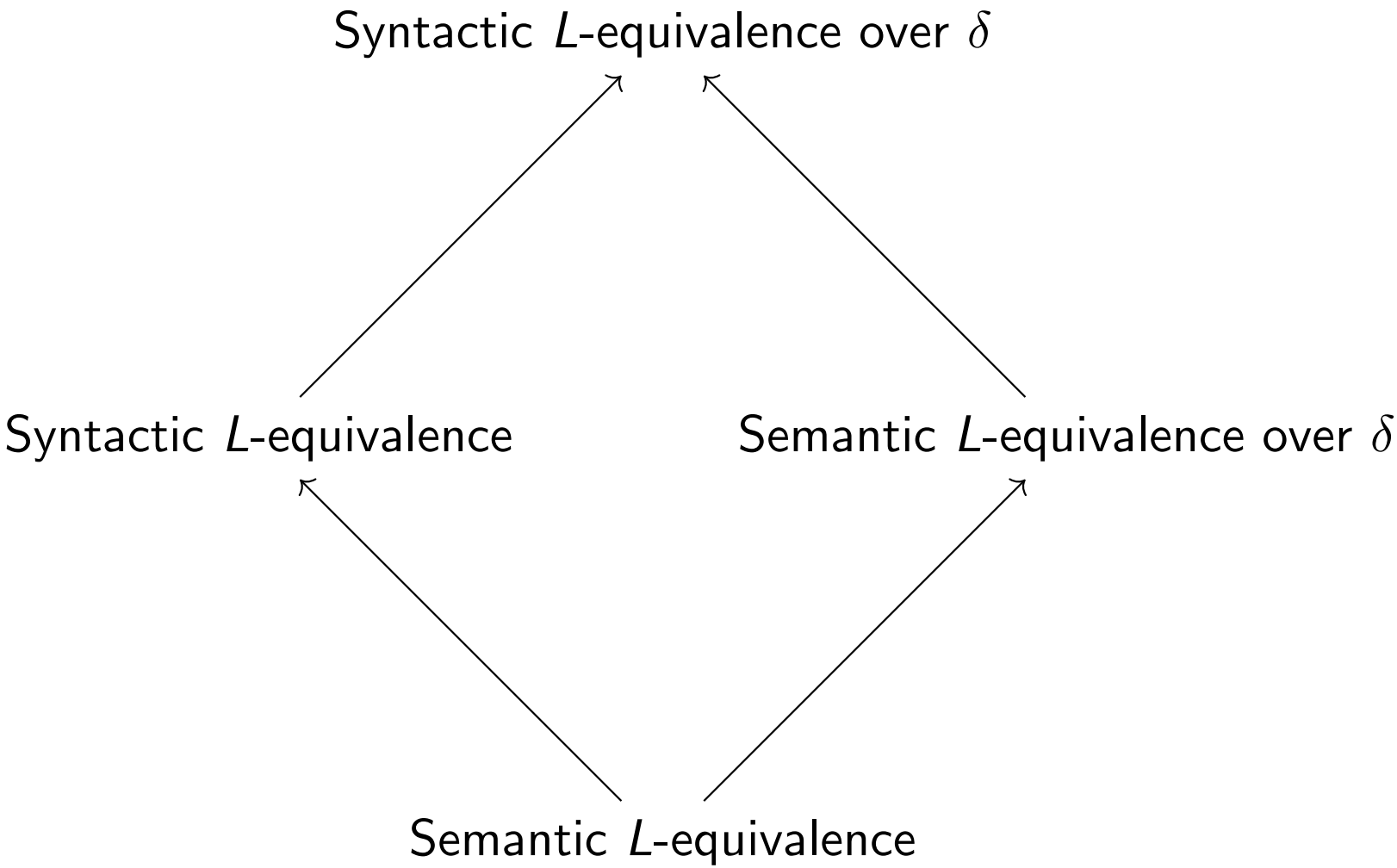

Figure: The relation of entailment between the four notions of empirical equivalence.

Syntactic and semantic equivalence

Two theories are

- *syntactically L-equivalent (over δ)* just in case they entail the same (δ-relativized) *L*-sentences, and
- *semantically L-equivalent (over δ)* just in case the models satisfying them have the same (δ-restricted) *L*-reducts.

Theoretical equivalence

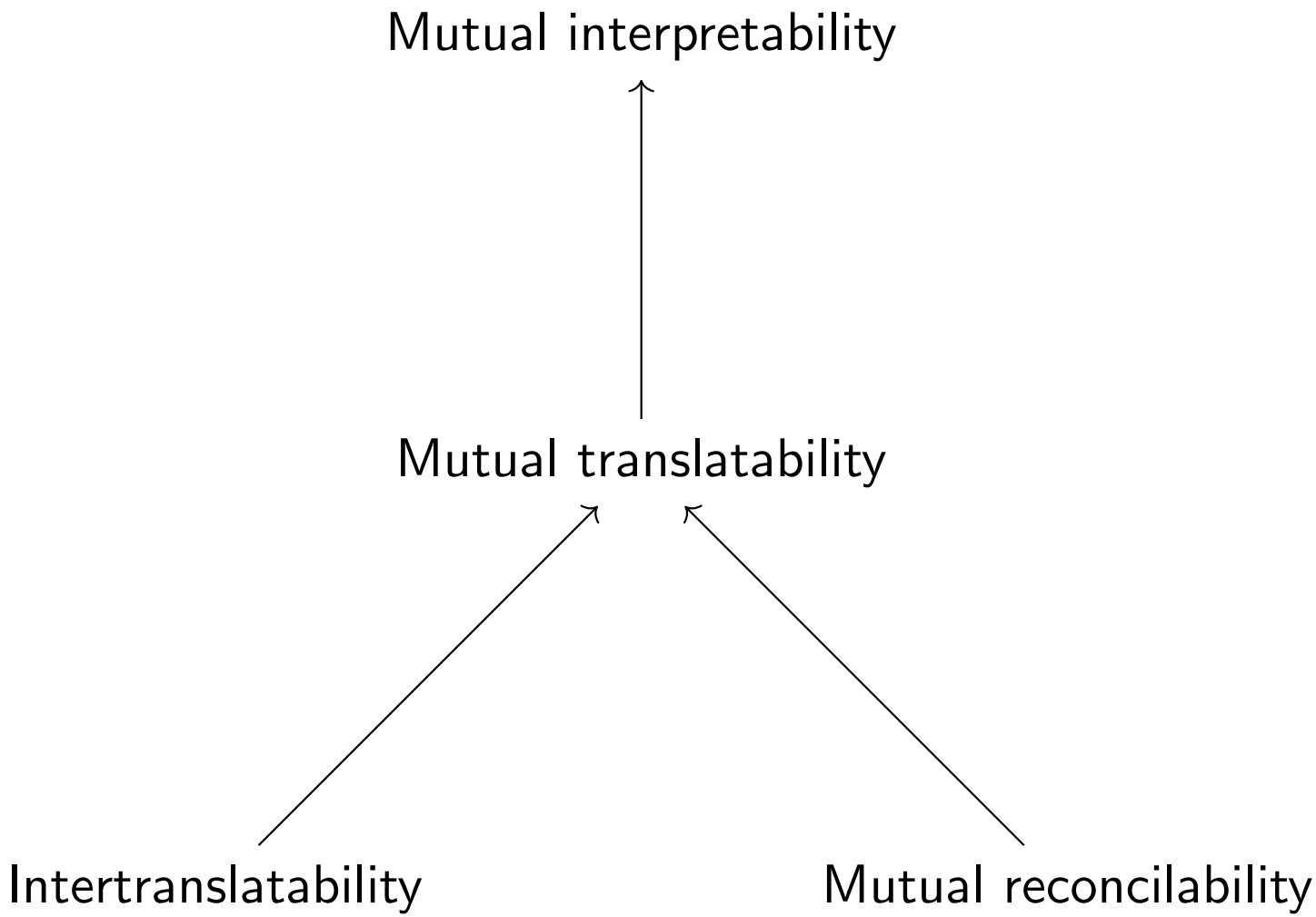

Figure: The relation of entailment between the four notions of theoretical equivalence.

Interpretability

Definition

An *interpretation* is a function I from L_1-formulas to L_2-formulas such that $I(x = y)$ is $x = y$ and, for any n-place L_1-predicate P, there is an L_2-formula $\varphi(x_1, ..., x_n)$ such that $I(P\bar{x}) = \varphi(\bar{x})$, and there is an L_2-formula $\delta(x)$ (a so-called *domain formula*) such that, for any L_1-formulas φ and ψ,

(i) $I(\neg\varphi) = \neg I(\varphi)$

(ii) $I(\varphi \to \psi) = I(\varphi) \to I(\psi)$

(iii) $I(\forall x\varphi) = \forall x(\delta(x) \to I(\varphi))$

Definition

An L_1-theory T_1 is *interpretable* by an L_2-theory T_2 just in case there is an interpretation I from L_1-formulas to L_2-formulas such that, for any L_1-sentence φ, if $T_1 \vdash \varphi$ then $T_2 \vdash I(\varphi)$. Relative to I, we then say that T_2 *interprets* T_1.

A non-interpretability lemma

Lemma (Feferman)

Let T be a consistent theory, let I be an interpretation such that $I(PA) \subseteq T$, and assume that $\alpha(x)$ is a Σ_1-formula representing T in PA. Then T cannot interpret $T \cup \{I(Con_\alpha)\}$.

S. Feferman. Arithmetization of metamathematics in a general setting.
Journal of Symbolic Logic, 31(2):269–270, 1966

My main result

Theorem

Let T be a theory in vocabulary L_T, and let $L \subseteq L_T$. Assume that (i) T is consistent, (ii) T does not have any finite models, and that (iii) there is a recursive theory T^* semantically L-equivalent to T such that T and T^* are jointly inconsistent. If T^* can interpret T, there is a finite extension T' of T^* such that (iv) T and T' are semantically L-equivalent, and (v) T cannot interpret T'.

Answering Quine's question

- Given a theory T such that $T \vdash \exists \bar{x} P \bar{x}$ for some theoretical predicate P, constructing an empirically equivalent rival T^* is a trivial matter: just replace P everywhere with a new predicate P^*, and add the sentence $\neg \exists \bar{x} P \bar{x}$.
- But the two theories are mutually translatable: we can translate every theorem of T to a theorem of T^* by replacing P with P^*, and we can translate every theorem of T^* to a theorem of T by replacing P^* with P and $P\bar{x}$ with $\neg \forall x (x = x)$.
- If T is consistent, recursive, and does not have any finite models, then T^* will inherit these properties.
- In that case, as a consequence of my main result, one can extend T^* with a single sentence (saying, essentially, that T^* is consistent), thereby producing a theory with the same empirical content as T^*, but one that T *cannot* interpret.