DESCRIPTION FOR THE GENERAL PUBLIC

One of the most fundamental achievements of logic is the axiomatisation of substantial parts of mathematics. This has been done by introducing intuitively obvious principles and rules which allow us to derive known theorems of algebra, analysis, geometry or other canonical parts of mathematics. Such axiomatic systems, that are sufficiently powerful to formally represent and develop large domains of mathematical reasoning, are called *foundational theories*. Their investigation is a major focus of logic.

Foundational theories can describe properties of syntactic objects, such as strings of characters from a finite alphabet (words), strings of words (sentences) or strings of sentences (proofs). Moreover, foundational theories can define which sentences are their axioms. Since proofs in foundational theories use rules from some well-defined collection, we can decide whether a given string of sentences satisfies these rules. In such a way, we can check whether that string is a well-formed proof in the given theory which allows foundational theories to tell which sentences are their theorems. Thus, foundational theories are ‘reflective’ in the sense that they can reason about themselves.

Foundational theories can define which sequences of sentences are correct proofs *from the axioms of the given theory*. This allows to define a theorem as a sentence such that it has a proof. Thus we can ask, whether a given theory proves that some given sentences are its theorems while some other (say, overtly absurd) are not. It turns out that *no* foundational theory can show that an explicit absurdity is underivable. Moreover, every theory is subject to this fundamental limitation. This result, known as Gödel's Second Theorem, is one of the most remarkable achievements of logic or perhaps even of the entirety of science.

Although no foundational theory can verify that an absurdity is underivable, one can argue that by accepting a given theory, at the same time we have to assume that this very theory is consistent and consequently can derive no inconsistency. However, the consistency of a given theory cannot be proved as one of its theorems, that is, it is not an *explicit commitment* of the theory. In effect, one can only say that the consistency statement is a consequence of the theory in a certain broader sense. Namely, the rational agent cannot accept a theory and at the same time hold that it is inconsistent. The acceptance of the consistency statement is not a matter of our decision or accepted conventions. In such a case, we say that the consistency statement is an *implicit commitment.*

Classical works of Solomon Feferman and Alan Turing answered the question of what else can be proved under the assumption that accepting a given theory entails accepting its consistency statement or related principles. It turned out that this method of extending a given theory yields the same result as several other natural procedures which seem unrelated at the first sight.

The aforementioned works described what happens if we assume that a concrete kind of implicit commitments (consistency statements) is represented as a procedure of adding new axioms. In our project, we try to understand the general mechanisms of creating implicit commitments of foundational theories. In other words, we try to describe in what precise and mathematically sound sense, by accepting a given theory we should accept its consistency and what are other implicit commitments of foundational theories.

Besides that, in our project we plan to investigate other possible kinds of implicit commitments of theories which are not as well-understood as consistency statements and related principles. The basic example are truth-theoretic statements. It seems that if we accept a theory, we should also accept that the axioms of that theory are true. However, in order to even express that the axioms of a given theory are true, we should in the first place formulate some further assumptions governing the behaviour of the notion of truth.

One example of such a truth-theoretic assumption is compositionality, which is the principle that the truth value of a sentence depends on the truth values of its components. For instance, if a sentence *A* is true and a sentence *B* is true, then the sentence *A and B* should also be true. There are many other similar principles investigated in the context of truth theories. However, their strength is often unclear. It should be emphasised that by 'strength' one can mean several different, precisely defined notions. Nevertheless, their common thrust is that a stronger or more committing theory (in our case, a theory obtained by adding a truth predicate) will 'say more' than another theory (in our case, the base theory before adding the truth predicate). In the project we try to understand how strong the implicit commitments of truth principles really are.