**POLITECNICO DI MILANO**

SCHOOL OF INDUSTRIAL AND INFORMATION ENGINEERING
Master of Science in Mathematical Engineering

# Gödel's Disjunction and Artificial Intelligence

Supervisor: Prof. Giovanni Valente

Master thesis of:
Enrico Bertino,
matricola 850375

Academic Year 2018-2019

# Abstract

This thesis stems from a personal interest in the world of artificial intelligence (AI). Working in the practical part of this branch of computer science, I realized that AI is often misunderstood from both people and the scientific community. So I started to explore the debates about AI that are underway today, both from a mathematical and a philosophical point of view.

The scope of the thesis is an overview of Godel's Incompleteness Theorems and how they are used within the debate on the mechanism of the human mind. In this context, it is essential to talk about the "Godel's disjunction" according to which either the power of the human mind cannot be expressed by any finite machine or absolutely unsolvable problems exist.

## Sommario

Questa tesi nasce da un interesse personale al mondo dell'intelligenza artificiale (AI). Lavorando nella parte applicativa di questa branca della computer science, mi sono accorto di come l'AI sia mal compresa dalle persone, sia nel mondo quotidiano ma spesso anche all'interno della comunità scientifica, e questo mi ha portato a interrogarmi sui fondamentali dell'AI e a ricercare dibattiti in corso sia dal punto di vista matematico che filosofico.

L'obiettivo della tesi è fare una panoramica sui Teoremi di Incompletezza di Godel e su come essi siano usati all'interno del dibattito sul meccanicismo della mente umana. A tale scopo, viene riportata e analizzata la cosiddetta "Disgiunzione di Godel", disgiunzione secondo la quale o il potere della mente umana non può essere espresso da nessuna macchina finita o esistono problemi assolutamente irrisolvibili dall'uomo.

# Contents

# Introduction

This thesis stems from a personal interest in the world of artificial intelligence (AI) and in particular on the way this subject has been discussed within the scientific and philosophical community.

AI is not a new term, there has been conversation about AI since ancient times and the mathematical community started to formalize it in the first half of the 20th century. However only in the last two decades AI has become one of the most popular scientific topics. We can attribute the merit to scientific advances both in terms of hardware (increase of computational power), software (increase of parallelization capacity) and modeling (neural networks). It seems that the so-called "AI winter" is over and the results achieved have started to have a real impact on the world where we live. The term AI is on everyone's lips regardless of extraction, culture or location and there have been events with a high media attention, such as when in 2007 IBM Watson beat the two world champions at "Jeopardy!" or when in 2016 Google DeepMind's AlphaGo beat the strongest players of Go (which is considered one of the most complex board games in the world).

Most of the time, however, the term AI is used improperly and it is difficult to realize what this field really covers. Perhaps because of media events, or science fiction movies or perhaps because of the aggressive marketing undertaken by technology companies, there are too many cases in which AI is associated simply with magic or with robots that will conquer the world.

Today, many companies in the world are investing large amounts of capital in AI, but what kind of AI is it? And what do we mean with AI? Perhaps the definition problem stems from the intrinsic difficulty of defining a real discipline. Russell and Norvig [2002] in their book AIMA (cornerstone of modern artificial intelligence), try to solve this definition problem characterizing AI with its goal. The definition should therefore be of the form "AI is the field that aims at building...". S. J. Russell, for example, describes AI as a field dedicated to creating intelligent agents that work by taking tuples of perceptions from the surrounding environment and reproducing behavior based on these perceptions. These agents are implemented by a program on a machine and try to maximize the expected value of a utility function. We can see how the definition is set in terms of optimization of a utility function, in the sense that the agent, given a task that has a goal, tries to find the best result for him. It is clear that this type of definition aims to focus only on the result of a single human action. This is an appropriate definition for all the AI applications we have available today, falling in the so-called "Weak" AI. We can in fact distinguish between "Strong" and "Weak" AI by taking note of the different goals that these two versions of AI strive to reach.

> **Strong AI** (or General AI) seeks to create artificial persons: machines that have all the mental powers we have, including phenomenal consciousness.
> **Weak AI**, on the other hand, seeks to build information-processing machines that appear to have the full mental repertoire of human persons [Searle, 1997].

For the Strong AI, I prefer the definition given in 1973 by Newell, one of the precursors of the modern AI:

> AI is the field devoted to building artifacts that are intelligent, where 'intelligent' is operationalized through intelligence tests, and other tests of mental ability (including, e.g., tests of mechanical ability, creativity, and so on).

With this definition, we leave the information technology field and we enter a more philosophical field in which we consider intelligence as a set of human abilities and not as the optimization of a specific utility function.

Trying to bring the discussion to a more formal level, we can transfer our questions from the computer world to the world of idealized machines (Turing machines) and again to formal systems (Chapter 2). At this point we can ask ourselves: have there been attempts to prove mathematically that the capabilities of the human mind exceed any mechanism or formal system? Many attempts lead to one person: Kurt Gödel.

It is fair to say that there is no mathematical theorem that has aroused as much interest among both mathematicians and non-mathematicians as Gödel's Incompleteness Theorem, which appeared in 1931. The popular impact that this theorem has had in recent decades can be seen in every field. Unfortunately, many references to the incompleteness theorem outside the field of formal logic are based

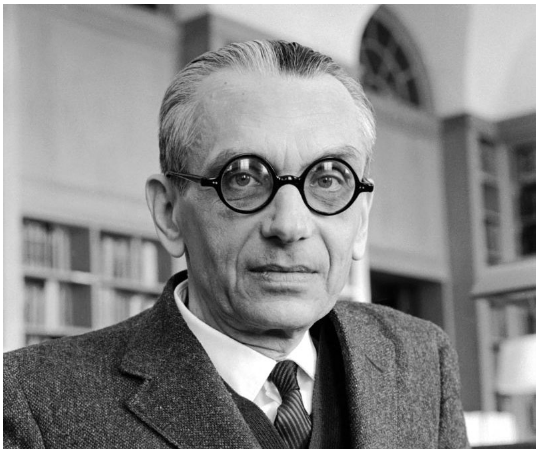

Figure 1: Kurt Gödel

on gross misunderstandings. Sokal and Bricmont [1998] and Franzén [2005], point out that "Gödel's theorem is an inexhaustible source of intellectual abuses". So the question is: why are these theorems so cited and abused even today? We can probably find the reason for this interest in the fact that the theorems start from a concept that is very easy to understand from everyone: the liar paradox, which says "this sentence is false". It's fascinating because of its use of the self-reference for making the truth of the sentence indefinable. In fact, if this sentence were true then it would be in contradiction with what it says, that is to be false, and if instead it were false, then it would say to be true contradicting itself also in this case. This ease of first understanding leads people to approach the incompleteness theorems, ignoring that they are based on a formalism and a complexity that has been difficult to understand even for the mathematical community itself. Gödel's genius was to bring that *simple* concept back into an unassailable mathematical formalism, proving the incompleteness of arithmetic and shaking the entire mathematical community.

Some years later, other logicians and philosophers started to debate on different issues based on the Gödel's theorems, such as the mechanism of the human mind. This is a debate that has lasted for almost a century and that struggled to find the standards of mathematical rigor shared by the community. In recent years much effort has been spent to try to formalize this debate: this led to define the so-called "Gödel's disjunction":

> **either** ... the human mind (even within the realm of pure mathematics)
> infinitely surpasses the power of any finite machine
> **or** else there exist absolutely unsolvable diophantine problems.

There have been repeated attempts to apply Gödel's theorems to demonstrate that the powers of the human mind outrun any mechanism or formal system. Let us call it *Anti-Mechanism.* Consequently, *Mechanism* holds the possibility that there can exist a machine that has the same human cognitive capabilities.

Within this thesis, I will examine this disjunction by trying to make explicit the mathematical and philosophical assumptions. An interesting aspect of the disjunction is that it stays halfway between mathematical logic and philosophy, probably the main reason why the debate is still so undefined. The statement itself is of a philosophical nature since it includes informal concepts such as mind, machine, or absolutely unsolvable problem. Nevertheless, all the authors considered are substantially in agreement on the validity of the disjunction and on the fact that Gödel has established beyond any reasonable doubt that the problem of the mechanization of the human mind and of the existence of humanly unsolvable problems are linked and mutually dependent. In the light of Gödel's disjunction, those who want to argue in favor of the existence of humanly unsolvable problems could then rely on arguments in favor of the possibility of mechanizing the human mind. On the contrary, those who wish to argue against the Mechanism, could instead start from the fact that all problems are humanly solvable.

I will present the Incompleteness Theorems in Chap. 1, followed by the debate on Mechanism of the human mind in Chap. 2 and the Gödel's Disjunction in Chap. 3,4 and 5.

# Chapter 1

# Gödel's Incompleteness Theorems

In 1931, Kurt Gödel published his First and Second Incompleteness Theorems (or simply Gödel's Theorems). These Incompleteness Theorems settled some of the crucial questions of the day concerning the foundations of mathematics. in the early 1900s Hilbert, after the crisis in the foundations of mathematics, published a list of twenty-three unsolved problems in mathematics. The second one was about proving that the axioms of arithmetic are consistent. Gödel tried to solve this problem but succeeded in proving the opposite, shaking up the whole mathematical community. The theorems remain of the greatest significance for the philosophy of mathematics and it has also frequently been claimed that Gödel's Theorems have a much wider impact on very general issues about language, truth and the mind.

Gödel presented and proved his incompleteness theorem in an Austrian scientific journal in 1931. The title of his paper was "On formally undecidable propositions of Principia Mathematica and related systems I." Principia Mathematica (PM) was a work in three volumes by B. Russell and A. N. Whitehead, published 1910-1913, putting forward a logical foundation for mathematics in the form of a system of axioms and rules of reasoning within which all of the mathematics known at the time could be formulated and proved.

The *first incompleteness theorem* established that on the assumption that the system of PM satisfies a property that Gödel named $\omega$-consistency, it is incomplete, meaning that there is a statement in the language of the system that can be neither proved nor disproved in the system. Such a statement is said to be *undecidable* in the system.

The *second incompleteness theorem* showed that if the system is consistent, meaning that there is no statement in the language of the system that can be both proved and disproved in the system, the *consistency of the system* cannot be established within the system.

Although Gödel used in his proof the property of $\omega$-consistency, which is a stronger property than consistency, J. Barkley Rosser showed in 1936 that Gödel's theorem could be strengthened so that only the assumption of plain consistency is needed to conclude that the system is incomplete. Also, it was immediately clear that his result also applied to a wide range of axiomatic systems for mathematics. Today the incompleteness theorem is often formulated as a theorem about any formal system within which a *certain amount of elementary arithmetic can be expressed* and some basic rules of arithmetic can be proved (see Section 1.2 for more details).

In my opinion, the most fascinating aspect about these theorems and their resonance is that they were also discussed in a non-formalized context, despite their mathematical complexity. In fact, "consistent," "inconsistent," "complete," "incomplete," and "system" are words used not only as technical terms in logic, but in many different ways in ordinary language, so it is not surprising that Gödel's theorem has been considered by the most different people and has been associated with various ideas in some informal sense. On the other hand the kind of reasoning put forward in Gödel's paper was at the time unfamiliar to logicians and mathematicians, and even some accomplished mathematicians (for example, the founder of axiomatic set theory, Ernst Zermelo) had difficulty grasping the proof. This is perhaps the reason why it took twenty years for the theorems to be universally accepted and why even today there is a strong focus on debates on them.

In the Gödel disjunction analyzed in the next chapters, the two theorems will be used in the proofs of the different arguments. In particular, the first one is mainly used by Lucas and Penrose for proving the Anti-Mechanism and the second by Godel for his Disjunction. I will outline in this chapter the most relevant aspects of the theorems summarizing the works of Murawski [1999], Raatikainen [2005], Franzén [2005], Smith [2007] and mostly Raatikainen [2018b].

## 1.1 Basics

Here some preliminaries about arithmetic and logic that we need in this work.

**Numeral** The formal term ("numeral") canonically denoting the natural number **n** is abbreviated as $\underline{n}$. In the standard language of arithmetic used here, the number **n** is denoted by the term $0^{\prime\cdots\prime}$, where the successor symbol "$'$" is iterated n times.

That is, numerals which name 1, 2, 3, . . . are $0'$, $0''$, $0'''$, . . . .

**Quantifiers** In logic, *quantification* specifies the quantity of specimens in the domain of discourse that satisfy an open formula. The two most common *quantifiers* are "for all" $\forall$ and "there exists" $\exists$. *Bounded* quantifiers are often included in a formal language in addition to the standard quantifiers "$\forall$" and "$\exists$". In Peano arithmetic, there are two types of bounded quantifiers: $\forall n < t$ and $\exists n < t$.

A formula $A$ in first order language is quantifier-free if and only if it contains no unbounded quantifiers.

A theory has quantifier elimination if for every formula $A$, there exists another formula $A_{QF}$ without quantifiers that is logically equivalent to it.

**Syntactic consequence** A formula $A$ is a syntactic consequence of a set $\Gamma$ of formulas within some formal system $F$ if there is a formal proof of $A$ in $F$ from the set $\Gamma$.

$$\Gamma \vdash_F A$$

Syntactic consequence does not depend on any interpretation of the formal system.

**Semantic consequence** A formula $A$ is a semantic consequence within some formal system $F$ of a set of formulas $\Gamma$

$$\Gamma \models_F A,$$

if and only if there is no model $\mathcal{I}$ of $F$ in which all members of $\Gamma$ are true and $A$ is false. Or, in other words, the set of the interpretations that make all members of $\Gamma$ true is a subset of the set of the interpretations that make $A$ true.

**Recursive sets and recursively enumerable sets** First, there may be a mechanical method which decides whether any given number belongs to the set at issue or not (in which case the set is called "decidable" or "recursive"), and, second, there may be a mechanical method which generates or lists the elements of the set, number by number. In the latter case, the set is called "recursively enumerable" , i.e. it can be effectively generated or it is "semi-decidable". It is a fundamental result of the theory of computability that there are semi-decidable sets but are not decidable (i.e., not recursive).

**Consistency, Soundness, and Completeness** Very informally, consistency states that the system does not entail a contradiction; a soundness theorem for a deductive system expresses that all provable sentences are true; completeness states that all true sentences are provable. This concepts will be ananlyzed more formally afterwards.

**$\omega$-consistency** In his original proof, Gödel used his specific notion of $\omega$-consistency, and for some purposes, it is still convenient to follow Gödel's original approach. Given a free variable $x$, a formalized theory $F$ is $\omega$-consistent if it is not the case that for some formula $A(x)$, both $F \vdash \neg A(\underline{n})$ for all $\mathbf{n}$, and $F \vdash \exists x A(x)$. Naturally this implies normal consistency.

**Existential formulas and 1-consistency** Actually, a simple special case of $\omega$-consistency suffices in the first theorem; namely, the assumption is only needed with respect to what logicians call $\sum_1^0$-formulas; these are, roughly, the purely existential formulas; more exactly, formulas of the form $\exists x_1 \exists x_2 \ldots \exists x_n A$, where A does not contain any unbounded quantifiers. This restricted $\omega$-consistency is called 1-consistency.

1-consistency can be expressed intuitively simply as the requirement that the formal system in question does not prove any false $\sum_1^0$-sentences (i.e., the system is sound at least in the case of such sentences).

**Universal formulas** A universal $\prod_1^0$-formulas is a formula of the form $\forall x_1 \forall x_2 \cdots \forall x_n A$ where $A$ is a quantifier-free formula.

## 1.2 Arithmetical Theories

In the statements of the incompleteness theorems, there is the requirement that "a certain amount of elementary arithmetic can be carried out". Let us see what it means and for what theories we can apply the Incompleteness Theorems.

**Arithmetical Theories** The weakest standard system of arithmetic that is usually considered in connection with incompleteness and undecidability is so-called *Robinson arithmetic* (due to Raphael M. Robinson), standardly denoted as **Q**. As axioms, it has the following seven assumptions:

- $\neg\,(0 = x')$
- $x' = y' \rightarrow x = y$
- $\neg(x = 0) \rightarrow \overline{\exists} y\,(x = y')$
- $x + 0 = x$

- $x + y' = (x + y)'$
- $x \times 0 = 0$
- $x \times y' = (x \times y) + x$

Where "$x'$" is the successor function, + the addition and and $x$ the multiplication. 0 is the only constant and denotes the number zero. Adding to these elementary axioms the axiom scheme of induction:

$$\phi(0) \wedge \forall x \left[\phi(x) \to \phi\left(x'\right)\right] \to \forall x \phi(x)$$

results in (first order) *Peano Arithmetic* (**PA**). Note that unlike **Q**, **PA** contains infinitely many axioms, because all instances of the induction scheme (one corresponding to every formula $\phi(x)$ with at least one free variable of the language) are taken as axioms. **PA** is generally taken as the standard first-order system of arithmetic.

Another natural and much-studied arithmetical system, which lies between **Q** and **PA**, is *Primitive Recursive Arithmetic* (**PRA**). It contains not just the above axioms of **Q** governing successor, addition and multiplication, but also defining axioms for all *primitive recursive functions*, and the application of the induction scheme is restricted to quantifier-free formulas (i.e., $\phi(x)$ is not allowed to contain any (unbounded) quantifiers).

However, essentially the same system is obtained if one takes just the axioms of **Q** and the induction scheme restricted to purely existential formulas ($\sum_1^0$-formulas). Moreover, $\sum_1^0$-induction can be shown to be equivalent to the induction scheme restricted to purely universal formulas ($\prod_1^0$-formulas). **PRA** is sufficient for developing the theory of syntax for formalized theories [Raatikainen, 2018b].

To summarize: when it is said, in the context of the incompleteness theorems, that "*a certain amount of elementary arithmetic can be carried out*" in a system, this usually means that it contains **PRA** or at least **Q**. For the first incompleteness theorem, **Q** is sufficient; for the standard proofs of the second theorem, something like **PRA**, at a minimum, is needed [Raatikainen, 2018b].

## 1.3 Representability and Completeness

We also need the notion of representability of sets and relations in a formal system F. More precisely, two related notions are needed.

**Strong representation** A set $S$ of natural numbers is strongly representable in $F$ if there is a formula $A(x)$ of the language of F with one free variable $x$ such that for every natural number $n$:

$$\begin{aligned} \boldsymbol{n} \in S &\implies F \vdash A(\underline{n}) \\ \boldsymbol{n} \notin S &\implies F \vdash \neg A(\underline{n}) \end{aligned}$$

**Weak representation** A set $S$ of natural numbers is weakly representable in $F$ if there is a formula $A(x)$ of the language of $F$ such that for every natural number $n$:

$$\boldsymbol{n} \in S \Leftrightarrow F \vdash A(\underline{n})$$

As the incompleteness results in particular teach us, there are sets which are only weakly but not strongly representable (the key example being the set of statements provable in the system).

In the case of both kinds of representability (weak and strong), there is always a simple existential $\sum_1^0$-formula, which represents the set in question, and usually such a formula is used to represent $S$.

Quite independently of the particular formal system chosen, exactly the decidable, or recursive, sets are strongly representable, and exactly the semi-decidable, or recursively enumerable sets are weakly representable. This holds for all formalized systems which contain **Q**. Instead of using the notion of *representability*, Gödel took a different approach by speaking of sets being **decidable** in a formal system F (*entscheidungsdefinit*). If the proofs of $F$ are systematically generated, it will be eventually determined, for any given number $n$ - whether it belongs to $S$ or not - given that $S$ is strongly representable in F.

In sum, we have:

**The Representability Theorem**
In any consistent formal system which contains **Q**:

1. A set (or relation) is strongly representable if and only if it is recursive;
2. A set (or relation) is weakly representable if and only if it is recursively enumerable.

## 1.4 Arithmetization of the formal language: Gödel's numbering

Godel needed to bring the concepts of any formal language into arithmetic. So he took the language of a formal system, which is always precisely defined, and fixed a correspondence between the expressions of that language and the system of natural numbers. A coding, *arithmetization*, or *Gödel numbering*, of the language. The essential point is that the chosen mapping is effective: it is always possible to pass, purely mechanically, from an expression to its code number, and from a number to the corresponding expression. This was one of the most ingenious elements of Gödel that allowed him to apply concepts of common language such as the liar paradox to arithmetic.

One proceeds as follows: first, the primitive symbols of the language are paired with distinct natural numbers, "symbol numbers". A little number theory then suffices to code sequences of numbers by single numbers. Consequently, well-formed formulas, as sequences of primitive symbols, are each assigned a unique number. Finally derivations, or proofs, of the system, being sequences of formulas, are arithmetized, and are also assigned specific numbers. Such a code, the "Gödel number" of a formula $A$, is denoted as $\ulcorner A \urcorner$, and similarly for derivations.

In this way, syntactical properties, relations and operations are reflected in arithmetic: for example, $neg(x)$ is the arithmetical function that sends the Gödel number of a formula to the Gödel number of its negation; in other words,

$$neg(\ulcorner A \urcorner) = \ulcorner \neg A \urcorner$$

similarly, $impl(x, y)$ is the function which maps the Gödel numbers of a pair of formulas to the Gödel number of the implication of the formulas:

$$impl(\ulcorner A \urcorner, \ulcorner B \urcorner) = \ulcorner A \to B \urcorner$$

and so on. There is an arithmetical formula, call it $Fmla(x)$, which is true of $\boldsymbol{n}$ iff $\boldsymbol{n}$ is a Gödel number of a well-formed formula of the system. There is also an arithmetical formula $M(x, y, z)$ which is true exactly if one has a valid application of the rule of inference. In this way, *all the syntactic properties and operations can be simulated at the level of numbers, and moreover they are strongly representable in all theories which contain* $\boldsymbol{Q}$.

The same can be applied to proofs and the provability itself. As it is decidable whether a given sequence of formulas constitutes a proof of a given sentence, the binary relation "x is (the Gödel number of) a proof of the formula (with the Gödel number) y" can be strongly represented in all systems containing $Q$. Let us denote the formula which strongly represents this relation in $F$ itself as $Prf_F(x, y)$. The property of being provable in F can then be defined as $\exists x Prf_F(x, y)$. Let us abbreviate this formalized provability predicate as $Prov_F(x)$. It follows that the latter is weakly representable:

$$F \vdash A \Rightarrow F \vdash Prov_F(\ulcorner A \urcorner)$$

It is always possible to choose the provability predicate $Prov_F(x)$ to be a $\sum_1^0$ -formula.

## 1.5 Self-reference: the diagonalization lemma

As we said, Gödel's numbering was the key to apply common concepts like the self-reference to arithmetic. Gödel formalized it in the diagonalization lemma.

**The Diagonalization Lemma**
Let $A(x)$ be an arbitrary formula of the language of $F$ with only one free variable. Then a sentence $D$ can be mechanically constructed such that

$$F \vdash D \leftrightarrow A\left(\ulcorner D \urcorner\right)$$

In the literature, this lemma is sometimes also called "the self-referential lemma" or "the fixed point lemma". It has many important applications beyond the incompleteness theorems.

It is often said that given a property denoted by $A(x)$, the sentence D is a self-referential sentence which "says of itself" that it has the property $A$. Note that the lemma only provides a (provable) material equivalence between $D$ and $A(\ulcorner D \urcorner)$ (which states that both sides must have the same truth-value) and does not claim any sort of sameness of meaning.

## 1.6 First Incompleteness Theorems

We can now claim and prove the first theorem.

**Gödel's First Incompleteness Theorem**
Assume $F$ is a formalized system which contains **Q**. Then a sentence

$G_F$ of the language of $F$ can be mechanically constructed from $F$ such that:

- If $F$ is consistent, then F $\nvdash G_F$ .
- If $F$ is 1-consistent, then F $\nvdash \neg G_F$ .

Such an independent, or undecidable (neither provable nor refutable in $F$) statement $G_F$ in $F$ is often called *the Gödel sentence* of $F$.

In favourable circumstances, it can be shown that $G_F$ is true but unprovable. This is the case if, for example, the provability predicate $Prov_F(x)$ has been chosen as a $\sum_1^0$ -formula: The Gödel sentence is then provably equivalent to the universal formula $\forall x \neg Prf_F(x, \ulcorner G_F \urcorner)$. Such formulas can be proved false whenever they in fact are false: if false, there would be a number n such that $F \vdash Prf_F(n, \ulcorner G_F \urcorner)$ (this holds already in **Q**). This, however, would contradict the incompleteness theorem. Therefore, $G_F$ cannot be false, and must be true. For this reason, the Gödel sentence is often called *true but unprovable.*

Note that *Gödel's theorem* is the general incompleteness result of Gödel which concerns a large class of formal systems, while the *Gödel sentence* is the constructed, formally undecidable sentence which varies from one formal system to another. This is why it is important to include the subscript $F$ in $G_F$.

**Proof** The Diagonalization Lemma is applied to the negated provability predicate $\neg Prov_F(x)$. This gives a sentence $G_F$ such that:

$$F \vdash G_F \leftrightarrow \neg Prov_F(\ulcorner G_F \urcorner). \tag{G}$$

Thus, it can be shown, even inside F, that $G_F$ is true if and only if it is not provable in F. It is not difficult to show that $G_F$ is neither provable nor disprovable in F, if F only is 1-consistent.

For the **first half**, assume that $G_F$ were provable. Then, by the weak representability of provability-in-F by $Prov_F(x)$, F would also prove $Prov_F(\ulcorner G_F \urcorner)$. However, because F in fact also proves the equivalence (G), F would then prove $\neg G_F$ too. But this would mean that F is inconsistent. In sum, if F is consistent, then $G_F$ is not provable in F. For this first half, the assumption of the simple consistency of F suffices.

For the **second half**, it has to be assumed that F is 1-consistent (if $Prov_F(\ulcorner G_F \urcorner)$

has been chosen such that it is a $\sum_1^0$-sentence; otherwise, the more general assumption of $\omega$-consistency is needed).

Assume that $F \vdash \neg G_F$. Then F cannot prove $G_F$ , for otherwise F would be simply inconsistent. Hence no natural number **n** is the Gödel number of a proof of $G_F$, and because the proof relation is strongly representable, for all **n**, $F \vdash \neg Prf_F(\underline{n}, \ulcorner G_F \urcorner)$. If also $F \vdash \exists x Prf_F(x, \ulcorner G_F \urcorner)$, F is not 1-consistent, against the assumption. Therefore F does not prove $\exists x Prf_F(x, \ulcorner G_F \urcorner)$, in other words, by the definition of $Prov_F(x)$, F does not prove $Prov_F(\ulcorner G_F \urcorner)$. By the key equivalence (G), F also does not prove $\neg G_F$ (contradiction). □

## 1.7 Second Incompleteness Theorems

Given the arithmetized provability predicate, it is also easy to present an arithmetized consistency statement: pick some manifestly inconsistent formula (in arithmetical theories, a standard choice is $(0 = 1)$); let us denote it by $\psi$; (the arithmetized counterpart of) the consistency of the system can then be defined as $\neg Prov_F(\ulcorner \psi \urcorner)$. Let us abbreviate this formula by $Cons(F)$. The proof of the first part of the first incompleteness theorem can then presumably be formalized inside $F$. This gives:

$$F \vdash Cons(F) \rightarrow G_F$$

where $G_F$ is the Gödel sentence for $F$ provided by the first theorem. If $Cons(F)$ were provable in $F$, so would be $G_F$ , by simple logic. This would contradict Gödel's first theorem. Consequently, $Cons(F)$ cannot be provable in F either.

> **Gödel's second incompleteness theorem**
> Assume F is a consistent formalized system which contains elementary arithmetic. Then $F \nvdash Cons(F)$.

Which informally can be read as "If $F$ is a sound system, in which a certain amount of elementary arithmetic can be carried out, then its own consistency can not be proved in $F$".

There is a question of philosophical importance that should be mentioned here: as it stands, Gödel's second incompleteness theorem only establishes the unprovability of one sentence, $Cons(F)$. But does this sentence really express that F is consistent? Furthermore, might there not be other sentences which are provable and also express the consistency of F?

Giving a rigorous proof of the second theorem in a more general form that covers all such sentences has turned out to be very complicated and it will be omitted in this work. I just cite Feferman [2006] who said that it is customary to say that "whereas the first theorem and its relatives are extensional results, the second theorem is intensional: it must be possible to think that $Cons(F)$ in some sense expresses the consistency of F - that it really means that F is consistent."

# Chapter 2

# Towards the debate on Mechanism

In the 1930s, many logicians wondered about in what systems we could apply the incompleteness theorems. What they wanted to prove was the possibility of moving from a formal theoretical system to practical applications without losing the formalism necessary for the theorems. In other words the goal was to consider, instead of decidable sets or properties, computable functions which are useful in considering real applications, like machines. Gödel, Alonzo Church and Alan Turing have shown that these two concepts (decidable sets and computable functions) are interchangeable. They independently presented different proposals for an exact mathematical definition of computable functions and, consequently, of decidable sets. Recall that a set of axioms and the notion of provability are necessary for a formalized system to be decidable. Moreover, since the label ”recursive function” has, for historical reasons, been dominant in the logical literature, decidable sets are often called ”recursive sets”.

## 2.1 Turing machines and Church-Turing thesis

The purpose of Turing and Church (and, indirectly, Godel) was to find a precise definition of **effective procedure** (or ”mechanical procedure” or ”algorithm”), i.e. *processes that can be performed through a finite sequence of steps from an idealized agent starting from a finite number of instructions.* We can summarize the results that Gödel, Turing and Church independently found:

- In 1933, Kurt Gödel created a formal definition of the class of general recursive functions. The class of general recursive functions is the smallest class of functions which includes all the constant functions, the projections, the successor function, and which is closed under function composition, recursion, and minimization.

- In 1936, Alonzo Church created a method for defining functions called $\lambda$-calculus. Within $\lambda$-calculus, he defined an encoding of the natural numbers called the Church numerals. A function on the natural numbers is called $\lambda$-computable if the corresponding function on the Church numerals can be represented by a term of the $\lambda$-calculus.
- In 1936, before learning about Church's work, Alan Turing created a theoretical model for machines, now called Turing machines, which could perform calculations by manipulating symbols on a tape as input. Given an adequate encoding of natural numbers as symbol sequences, a function on natural numbers is called Turing-computable if a Turing machine can compute the corresponding function on encoded natural numbers.

Church [Church, 1936] and Turing [Turing, 1936–1937] showed that these three classes of computable functions, formally defined, coincide: a function is *lambda*-computable if and only if it is Turing-computable if and only if it is general recursive.

Among these, as Gödel [1986] pointed out, Turing's analysis of *fictitious and abstract computing machines (Turing machines)* was particularly relevant, as well as Church's work on the $\lambda$-calculus. The equation of this intuitive notion is often called "The Church-Turing thesis".

**Church-Turing thesis**
*The notion of Turing machine completely captures the concept of effective procedure.*

As it will be explained in Chapter 3, the Church-Turing thesis is one of the fundamental assumptions, together with the second incompleteness theorem, which is the basis of the Gödel's disjunction proof.

Another fundamental assumption is the fact that Turing machines in a certain sense constitute the mechanical counterpart of the theories studied by logicians. In fact, given a theory $F$ and an adequate numerical coding (like Gödel numbering) of the formulas in the language of this theory, it is possible to determine through an effective procedure whether a given number is the code of a formula of $F$. Since, therefore, the proofs are nothing more than finite sequences of formulas, it is still possible to determine whether a given number is the code of a proof and, when it is, we can actually determine the code of the last formula of the sequence, that is the proven theorem. It follows that, given a theory, one can define a mechanical procedure where one can control, for each natural number, if it is the code of a proof of $F$ and that, when it is, it produces the code of the theorem established by the proof. We can then associate to each theory a Turing machine that enumerates all the theorems of $F$.

Conversely, given a Turing machine $M$ and an appropriate formal language, $M$ can be made to correspond to an equivalent theory in the specified language, simply by choosing a suitable encoding so that all the numbers that $M$ can enumerate, are codes of formulas of the considered language and, finally, defining the theorems of $F$ as the deductive closure of the formulas enumerated by $M$ according to this encoding [Feferman, 2006]. We therefore have the following (meta) theorem which mathematically establishes the relation between Turing machines and logical theories.

> **Isomorphism theories-machines**
> Given any logical theory $F$, there is a Turing machine whose input and operating rules correspond to the axioms and rules of $F$ and whose output consists of all and only the theorems of $F$. Conversely, given any Turing machine $M$, there is a theory whose axioms and logical rules correspond to the inputs and rules of the machine and whose theorems are all and only the outputs of $M$.

It follows that we can talk about Turing machines the same way we talk about logical theories: everything that can be done through an automatic machine can be done through a corresponding logical theory and vice versa. So we can analyze the problem of Mechanism, i.e. the possibility of the mechanization of the human mind, as a logic problem. The isomorphism between formal systems and Turing machines also allows us, in the context of the discussion on the Disjunction, to ignore the formal definition of the Turing machine and to speak freely of "machine", i.e. any device capable of manipulating symbols and to prove syntactically true theorems [Beccuti, 2018].

We can also freely speak about the axioms or consistency of a Turing machine, always meaning the axioms and consistency of the theory corresponding to the machine considered on the basis of the aforementioned isomorphism. Similarly we can talk about the application of the second incompleteness theorem to Turing machines. Given a consistent Turing machine, there is a declaration (the one that expresses the consistency of the machine itself) that this machine cannot prove. Finally the Mechanistic thesis that we will discuss later can be expressed, taking into account the Church-Turing thesis and without losing its generality, as:

**Mechanistic thesis**
*The human mind in the field of pure mathematics is a Turing machine.*

Therefore, in this context, speaking of the mechanization of the mind means speaking of the possibility of simulating the capabilities of the mind through a machine.

## 2.2 Artificial Intelligence

The term 'artificial intelligence' made its advent at DARPA-sponsored summer conference at Dartmouth College, in Hanover, New Hampshire. Certainly the field of AI was in operation before 1956. For example, in a famous *Mind* paper of 1950, Alan Turing argues about the question "Can a machine think?" that he reformulated in the context of its Turing Test (TT) as "Can a machine be linguistically indistinguishable from a human?". The TT was a test in which a woman and a computer are sequestered in sealed rooms, and a human judge asks questions by "teletyping" to them. If the judge can do no better than 50/50 when delivering a verdict as to which room houses which player, we say that the computer in question has passed the TT. Turing predicted that his TT would be passed by 2000, but even today the most articulate of computers still can't meaningfully debate a sharp toddler. Moreover, while in certain focused areas machines out-perform minds (e.g. Deep Blue at chess, Watson at Jeopardy! or Alpha Go at Go), minds have a capacity for cultivating their expertise in virtually any sphere. AI simply hasn't managed to create general intelligence; it hasn't even managed to produce an artifact indicating that eventually it will create such a thing [Bringsjord and Govindarajulu, 2018].

But what is the definition of AI? AI is a field that can be studied from the perspective of many fields, like philosophy, mathematics and logic, computer science, biology, ... . Philosophers know better than anyone else that it can be extremely complex to give a definition that can satisfy all the different profiles that work in the field. Russell and Norvig [2002] in their book AIMA (cornerstone of modern artificial intelligence), characterize the definition of AI with its goal. The definition should therefore be of the form "AI is the field that aims at building...". All possible definitions can be included in the four areas listed below [Bringsjord and Govindarajulu, 2018]:

1. ... systems that think like humans
2. ... systems that think rationally
3. ... systems that act like humans

4. ... systems that act rationally

We can find many definitions in the literature. For example, philosopher Haugeland [1985] falls into the first category when he says that AI is:

> *The exciting new effort to make computers think ... machines with minds, in the full and literal sense.*

The second is instead defended by Winston [1992]. The third category is occupied most prominently by Turing, whose test is passed only by those systems able to act sufficiently like a human. Luger and Stubblefield [1993] seem to fall into the forth category when they write:

> *The branch of computer science that is concerned with the automation of intelligent behavior.*

Also Russell and Norvig themselves are firmly in the forth category. They describe AI as a field dedicated to creating intelligent agents that work by taking tuples of perceptions from the surrounding environment and reproducing behavior based on these perceptions. These agents are implemented by a program on a machine and try to maximize the expected value of a utility function.

We can see how these types of definitions are very far from each other: some are more philosophical, others concerns computer science and others biology. Only some of them treat the concept of the human mind. So how can we create a parallelism with Mechanism? We need to get more specific and to introduce a dichotomy in the definition: let's distinguish between "Weak" and "Strong" AI.

### 2.2.1 Weak AI

"Weak" AI is informally an AI system that can only act *like* it thinks and has a mind. Searle [1997] defined it as "information-processing machines that *appear* to have the full mental repertoire of human persons".

A Weak AI is an AI with no or limited ability to self-modify or generalize. For example, a machine that plays chess might have superhuman ability in the chess game, but it can only play chess. While it might tune its underlying model and slowly improve, it cannot modify itself in a deep enough way to generalize to other tasks. For this reason, sometimes Weak AI is also referred as "narrow" AI or "applied AI". In general, in the weak AI one do not care about human cognitive processes but exclusively about solving specific problems.

AI has achieved far greater commercial success and academic respectability by focusing on specific sub-problems where they can produce verifiable results and commercial applications, such as artificial neural networks, computer vision or data mining. These ”applied AI” systems are now used extensively throughout the technology industry, and research in this vein is very heavily funded in both academia and industry.

### 2.2.2 Strong AI

Searle [1997] defines ”Strong” AI as ”artificial persons: machines that have all the mental powers we have, including phenomenal consciousness”. Strong AI, also called Artificial general intelligence (AGI), has the capacity to understand and learn any intellectual task that a human being can, i.e. the capacity to perform the full range of human cognitive abilities. According to this definition, the machine should not be considered as an instrument but opportunely programmed so that it can be compared to the human mind, with an indistinguishable cognitive capacity. We can therefore create a parallelism between mechanism and Strong AI: studying the mechanization of the human mind is equivalent to study the possibility to create a Strong AI. In the following chapters, I will only refer to the term Mechanism in order to be aligned with the considered literature.

# Chapter 3

# Gödel's disjunction

In recent years some philosophers and logicians (to name a few, Feferman [2006], Fano and Graziani [2011], Stern [2018], Horsten and Welch [2016], Raatikainen [2018a]) have tried to put clarity on a debate that started some decade ago and that still remains undefined: what are the consequences that can be drawn from Gödel's incompleteness theorems about the human mind? Everything started from some Anti-Mechanistic arguments according to which the incompleteness theorems may indicate, or even prove, that the human mind surpasses any machine.

The main supporters of these arguments were Lucas and Penrose (Chap. 4). Their thesis was that the mathematical theorems that can be proven by an idealized human mind, can be generated by some effective procedure. This means, assuming the Turing-Church thesis, that the theorems that an idealized human mind can produce, are the output of a Turing machine.

Unlike Lucas and Penrose, Gödel did not believe that so strong thesis could be deduced as a direct consequence of its incompleteness theorems. Instead, he encapsulated his thought in a disjunction: one can believe in Anti-Mechanism only if he denies the possibility of the existence of humanly unsolvable problems. Officially he never took a part and he held the view that both disjuncts are consistent with incompleteness theorems.

The first time Godel spoke about this topic was in 1951 during a conference at the American Mathematical Society:

> it [second incompleteness theorem] makes it impossible that someone should set up a certain well-defined system of axioms and rules and consistently make the following assertion about it: All of these axioms and rules I perceive (with mathematical certitude) to be correct, and moreover I believe that they contain all of mathematics. If someone makes such a statement he contradicts himself. For if he perceives the

> axioms under consideration to be correct, he also perceives (with the same certainty) that they are consistent. Hence he has a mathematical insight not derivable from his axioms [Gödel, 1951] (page 309).

Does this mean that it is not possible to understand all mathematics in a single system of axioms? This can be considered the genesis of the Disjunction and it is clear that we need to discuss about the *interpretability* of this statement and the underlying *idealizations*.

## 3.1 The Disjunction from different perspectives

### 3.1.1 Subjective and objective mathematics

I want to start with the definition of "mathematics". During the aforementioned conference, Gödel introduced a distinction between subjective mathematics (the set of provable propositions starting from some system of axioms) and objective mathematics (the set of true propositions in an absolute sense). Do these sets coincide? If yes, then there is no hope of being able to understand all mathematics in a single axiomatic system, because if such a system exists, then the statement that expresses the consistence of the system cannot be provable in such a system, and this contradicts the initial assumption. If instead objective mathematics is distinct from subjective mathematics, then subjective mathematics could be liable to be understood in a single axiomatic system, but there would remain the problem of explaining the existence of true mathematical propositions that are not provable in a formal system. Gödel therefore supported the following disjunctive thesis: either subjective mathematics cannot be formalized, or objective mathematics is not reducible to subjective mathematics. In other words

> Either subjective mathematics surpasses the capability of all computers, or else objective mathematics surpasses subjective mathematics, or both alternatives may be true.

which he reformulated in more general terms:

> **Godel's Disjunction**
> So the following disjunctive conclusion is inevitable: Either mathematics is incompletable in this sense, that its evident axioms can never be comprised in a finite rule, that is to say, the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems

> ... (where the case that both terms of the disjunction are true is not excluded, so that there are, strictly speaking, three alternatives).. [Gödel, 1951] (Page 310).

### 3.1.2 Materialism and Platonism

Thus, Gödel's thesis is the following: either Mechanism is false or there are mathematical problems that we can not hope to solve. Let us try to make a parallelism with Materialism (or Physicalism), which holds that matter is the fundamental substance in nature and that all things, including mental states and consciousness, are results of material interactions. According to Gödel both alternatives of the Disjunction stand in clear opposition to Materialism in philosophy: of the **mind** in the case of the first disjunct, of **mathematics** in the case of the second one. He said:

> If the first alternative holds, this seems to imply that the working of the human mind cannot be reduced to the working of the brain, which to all appearance is a finite machine with a finite number of parts, namely, the neurons and their connections. [Gödel, 1951]

and the second one

> seems to disprove the view that mathematics is only our own creation; for the creator necessarily knows all properties of his creatures, because they can't have any others except those he has given to them. So this alternative seems to imply that mathematical objects and facts (or at least something in them) exist objectively and independently of our mental acts and decisions, that is to say, some form or other of Platonism or "realism" as to the mathematical objects [holds]. [Gödel, 1951]

Note that he intended Platonism as the existence of abstract objects, which are asserted to "exist" in a "third realm" distinct both from the sensible external world and from the internal world of consciousness.

If we accept the inferences and statements made so far, we have a variant of the Disjunction:

> either physicalism is false or else Platonism in mathematics is true, or both [Wang, 1996].

### 3.1.3 Psychology and Socrates

The relevance of the incompleteness theorems with regard to the human mind, although discussed by many logicians, remains however largely ignored by much of the contemporary philosophical and psychological debate, which does not seem to have understood its importance yet. As Benacerraf observed, in fact,

> it follows that [...] psychology as we know it is therefore impossible. For, if we are not at best Turing machines, then it is impossible, but if we are, then there are certain things we cannot know about ourselves or any others with the same output as ourselves. I won't take sides. But we can [...] reformulate the [philosophical] scope of Gödel's theorems as follows: if I am a Turing Machine then I am barred by my very nature from obeying Socrates' profound philosophical injunction: Know thyself. [Benacerraf, 1967]

## 3.2 Idealization

When a scientific model is studied in science philosophy, we often need to idealize some concepts or some hypothesis. In particular, idealization is the process by which models assume facts about the phenomenon being modeled that are not strictly real but make models easier to understand or solve. It is very important to indicate the specific idealizations underlying a debate in order to formalize the thesis and the arguments. Regarding the formulation of the Disjunction, some unanswered questions of epistemological, linguistic and idealistic nature persist. For example, the exact content of the Mechanistic thesis is not completely clear. The same applies to the content of Gödel's second disjunct on what human mathematicians can and cannot know (Gödel spoke also of "absolutely unknowable propositions").

Moreover, the relevance of completeness theorems for Mechanism depends on what the mechanist claims. The thesis that the human mind is, or can be modeled as, a computer or a Turing machine is often too vague to apply something as formal and precise as Gödel's theorems. The mechanist claims that there may be a machine whose results are the same as those of a human or a group of humans. But what kind of machine? What results? And what kind of human?

### 3.2.1 Unknowable arithmetic truths

In the second disjunct, Gödel comes to the conclusion of the possibility that arithmetic (absolutely) unknowable truths exist. If we stop on the first interpretation of the sentence, it might seem obvious that there are unknowable mathematical truths. Shapiro [2016] gives an intuitive example.

Let $p$ and $q$ be two prime numbers, each greater than $10^{1.000.000}$. If their product $n = pq$ were written in standard Arabic, decimal notation, it would have about two million digits. Consider the proposition $P$ for which $n$ has exactly two prime factors. By hypothesis, $P$ is true. However it is extremely unlikely that a human being could know $P$. That is, nobody can have convincing evidence for $P$. In fact, it is unlikely that anyone could even understand $P$ or even analyze the sentence. Thus, there are probably unknowable truths that can be plausibly stated. This example is somehow provocative, with the aim of showing how this part of the debate is philosophically unclear. In parallel, Penrose does not speak of unknowable truths but rather "unassailable". What does he mean? We could object that anyone can assail practically anything in a subjective way. Presumably, Penrose claims that propositions should have such strong evidence that no one could question them once the proof is understood. But some tests are so long and complicated that no human being can understand them. Therefore we need a formalization of the concepts of knowability and unassailable by the "human mind". This topic is well addressed by Koellner in [Horsten and Welch, 2016] which investigates the concepts of relative and absolute "provability" of arithmetic truths in the context of the second disjunct (see Chap. 5).

### 3.2.2 Absolutely undecidable truths

Other than *unknowable*, we can speak about *absolutly unprovable* (or *absolutely undecidable*). Hence, we should question about the concept of absolute provability: let us try to analyze who (and how) tried to define it.

An attempt has been made, starting from a proposal of Gödel himself, studying the properties of the absolute provability of mathematical propositions through a theory of arithmetic enriched with an *atomic modal operator of provability*, suitably axiomatized. In this direction, there are many contributions of Tharp and Fitch but their proposals are not entirely satisfactory since the propositions must necessarily contain the aforementioned modal operator and it is therefore a proposition that is only semi-mathematical, not an arithmetical statement as in Gödel's disjunction.

Another attempt, not entirely satisfactory, is that of Feferman and Solovay [1990] which highlight the existence of undecidable statements from the practical point of view, i.e. decisions *a priori* decidable but too complex to be decided in a reasonable time by a human being or a computer, like the statement "the value of the $10^{10^{10^{100}}}$-th decimal digit of $\pi$ is 7". This argument is valid if we are speaking about real humans and computers but it does not hold for the idealized ones.

Boolos [1982] shows instead that there are true "extremely undecidable" arithmetic statements, i.e. completely undecidable utterances characterized (within Peano arithmetic) by arithmetic-modal properties also possessed by all the other statements of Peano arithmetic and hence indistinguishable from them (within Peano arithmetic). Despite the suggestive name, Boolos does not prove that the extremely undecidable statements are unprovable in an absolute sense, but only inside Peano's arithmetic.

On the other hand, Williamson's approach is completely different. In 2016 he advances an original metaphysical argument for the conclusion that every true statement is absolutely provable based on the plausibility of the counterfactual existence of beings capable of proving it. Williamson, after having argued the non-analytic nature of mathematical truths, says that mathematical knowledge does not derive only from proofs, but, as far as the axioms are concerned, also from the evolutionary history of human beings, for which some statements seem "primitively convincing "by virtue of the specific hereditary characteristics of our brain [Williamson, 2016]. If we accept that every arithmetic statement is true or false and tertium non datur, then, argues Williamson, every arithmetic statement is absolutely provable (or refutable). Note, however, that Williamson does not believe that his argument can be used to support anti-mechanistic positions: even assuming that the aforementioned creatures are possible (future) human beings then for every arithmetic truth it is possible that there exists a mathematician who can prove it. This is different from saying that it is possible that there is a mathematician who can prove all the arithmetic truths.

Finally, further attempts are made in the direction of considering independent statements of set theory (for example, the axiom of choice or the hypothesis of the continuous in [Koellner, 2016a]) as possible candidates for being absolutely undecidable statements. As it was well explained by Koellner, this is Gödel's opinion until 1946, when he expresses the hope that we can reach a generalized completeness theorem for set theory based on the notion of Turing computability which establishes the impossibility of absolutely undecidable statements. Gödel then reaches the more mature view expressed at the 1951 conference. Koellner concludes that, in the context of set theory,

> There is at present no solid argument to the effect that a given statement is absolutely undecidable. We do not even have a clear scenario for how such an argument might go. [Koellner, 2006]

### 3.2.3 Human mind and Machines

In addition to the concept of demonstrability, we can ask ourselves, precisely, what is meant by "human mind". In the Disjunction, Godel speaks about "human mathematical mind". Does he mean the mind of a single mathematician or, more generally, the community of mathematicians?

Probably we are not talking about mathematicians as physical individuals, i.e. medical and physical limitations are not relevant. Presumably, the mechanist and the anti-mechanist are both talking about an ideal human, or a community of ideal humans (or mathematicians). Lucas and Penrose both refer to human capabilities "in principle". To idealize a mathematician, we should first understand how mathematicians know the mathematics they know. Then we have to somehow transfer this epistemology to the ideal humans, to understand how they get to know concepts and theorems. If we can clarify this point, we could then express the ideal mathematician's subjectivity at the center of Penrose's argument and we could define what he means by "unassailable" (see Chap. 4).

We can find an answer to the question about ideal humans starting from the concept of machine. As Shapiro [2016] noted, we must obviously idealize also on machines, whose idealization is however less complex than the human one. Like humans, today's digital computers have limitations, such as memory and materials, and are more like finite state machines. Computers are also subject to software malfunctions and bugs. It is in fact not sure that there can exist a *physical* computer whose output is one of the sets mentioned in the Church-Turing thesis. Could there be a hardware that corresponds to a human being in its arithmetic productions, reproducing both truths and errors? Or could there be hardware that matches the true arithmetic phrases of a given human? Maybe we know how to *produce* such a computer, but maybe we can't *build* it, depending on what we mean by *building.*

In any case, none of these considerations is relevant to the mechanistic thesis. On the machine side, the idealization solution is simple: we ignore the finite limits. In particular, suppose our machines never run out of memory, energy, time or work materials. Suppose then that they will never stop calculating just because they run out of material or memory or storage space: the traditional idea of potential infinity. Suppose further that the machines operate indefinitely without interruption, following their assigned programs perfectly. That is, by applying a known distinction between hardware and software, we ignore the hardware. We only consider what happens when programs are executed as they are wrote: we assume that the machines in question are abstract objects like Turing machines.

So in the debate we try to have idealizing hypotheses about humans analogous to those of Turing machines. We do not talk about the theorems that a subject produces, but about the theorems they can produce, an idealization similar to the one we invoke for Turing machines.

In short, both when we talk about the human and the machine, we talk about its potential without considering physical limits or malfunctions. Imagined creatures have unlimited lives, unlimited spans of attention, unlimited energy and unlimited materials at their disposal, just like Turing machines do. This let us to overcome the problems about unknowable truths like the aforementioned one on large prime numbers. However, it is assumed that these ideal mathematicians are like humans in every other aspect.

In chapter 5, I will report a more specific analysis on this theme, trying to formalize the concepts of idealized human mind and idealized machine through a parallelism with absolute and relative provability.

## 3.3 Fixing a framework for a proof

Given all these premises, one way to deal with the Disjunction is to define some hypotheses that we have to take for granted:

a. The second incompleteness theorem

b. The theories-machines isomorphism

c. The Church-Turing thesis

For some arguments, we need also (d) the consistency of the human mind, in order to apply the second incompleteness theorem to the human mind. Here a basic proof that uses the 4 hypotheses:

1. Assume that the human mind can be algorithmically simulated.
2. It follows that, for (c), there is a Turing machine capable of simulating the human mind in its ability to produce true theorems of arithmetic.
3. From (b) then follows that the mind can be simulated through a formal system of axioms.
4. Thus, if the mind is consistent, the second incompleteness theorem (a) applies to this system of axioms.
5. Thus there exists a true statement (d) that the system itself (the mind) cannot prove.

6. For the initial assumption, then, this statement is humanly (that is, absolutely) unprovable. □

Hence, if we suppose Mechanism, there are statements that are unprovable by any system.

Shapiro [2016] formulate a short more elegant proof of the Disjunction that uses the concepts of objective and subjective mathematics introduced previously, as well as the indefinable theorem of Tarski's truth:

1. Let $S$ be the set of mathematical truths recognizable as such by the human mind (subjective mathematics), and let $V$ be the set of mathematical truths (objective mathematics).

2. For definition, $S \subseteq V$.

3. Moreover, due to the Tarski's undefinability theorem, $V$ is not definable in the language of arithmetic and therefore a fortiori $V$ is not recursively enumerable.

D1 It follows that if $S = V$ then even $S$ is not recursively enumerable, and therefore Mechanism is false.

D2 Then it follows that, if Mechanism is true, $S \neq V$ and in such a case there exists a proposition $\phi \in V$ such that $\phi \notin S$. This means that $\phi$ is true but not humanly provable. □

**Reference to diophantine problems** It follows from some results of Gödel himself and from the joint work of Putnam, Robinson and Matiyasevich on the tenth Hilbert problem that the statement that expresses the consistency of a formal system is always (demonstrably) equivalent to a statement in the form

$$\forall x_1 \ldots x_n P(x_1 \ldots x_n) \neq 0$$

where P is a diophantine polynomial, i.e. a polynomial with coefficients and integer variables [Feferman, 2006]. In this sense in the statement of the Disjunction, Gödel speaks of "absolutely unresolvable diophantine problems".

# Chapter 4

# First Gödel's disjunct

In 1961, J.R. Lucas published ”Minds, Machines and Gödel” [Lucas, 1961], in which he formulated a controversial Anti-Mechanistic argument. The argument holds that Gödel's first incompleteness theorem shows that the human mind is not a Turing machine. The topic has generated many discussions that are still open due to the difficulty of defining a precise and formal perimeter in which to develop the debate. The reason of the interest around this debate is that the influential computational theory of the mind, which states that the human mind is a computer, is false if Lucas's argument is confirmed. That is, if Lucas's argument is correct, then ”strong artificial intelligence”, the idea that it is possible to build a machine that has the same cognitive capabilities as human beings, is false.

However, numerous objections to Lucas's argument have been presented. Some of these objections imply the consistency or inconsistency of the human mind: if we cannot establish that human minds are consistent, or if we can establish that they are actually inconsistent, then Lucas's argument fails. Others criticize various idealizations made by Lucas. Others claim that some parts of his argument are hardly defensible. Lucas's argument was reinvigorated when the physicist R. Penrose formulated and defended a similar argument in his two books, The Emperor's New Mind [Penrose, 1989] and Shadows of the Mind [Penrose, 1994]. Although there are similarities between the arguments of Lucas and Penrose, there are also some important differences. Penrose argues that the Gödelian arguments imply a series of statements concerning consciousness and quantum physics; for example, consciousness derives from quantum processes and may require a revolution in physics to obtain a scientific explanation. There have also been objections raised on Penrose's argument and on the various thesis that he deduces from it: some question the anti-mechanistic argument itself, while others question the solidity of its claims about consciousness and physics.

## 4.1 Lucas

Lucas used the proof of the first incompleteness theorem and tried to apply it to the human mind. First of all, consider a machine built to produce arithmetic theorems. Lucas starts from the *theories-machines isomorphism* to say that the operations of this machine are analogous to a formal system. Suppose now that we construct a Gödel sentence for this formal system. Since the Gödel sentence cannot be proved in the system, the machine will not be able to produce this sentence as a truth of arithmetic. However, a human can look at it and *see that the Gödel sentence is true.* In other words, there is at least one thing that a human mind can do that no machine can do. Therefore he says, "a machine cannot be a complete and adequate model of the mind" [Lucas, 1961]. In short, the human mind is not a machine.

Lucas [1990] describes his argument as follows:

> I do not offer a simple knock-down proof that minds are inherently better than machines, but a schema for constructing a disproof of any plausible mechanist thesis that might be proposed. The disproof depends on the particular mechanist thesis being maintained, and does not claim to show that the mind is uniformly better than the purported mechanist representation of it, but only that it is one respect better and therefore different. That is enough to refute that particular mechanist thesis.

Lucas therefore believes that a variant of his argument can be formulated to refute any future thesis of mechanists. In particular, imagine the following scenario:

- a mechanist formulates a particular mechanistic thesis claiming that the human mind is a Turing machine with a specific formal specification $S$.

- Lucas rejects this thesis by producing the Gödel sentence for $S$, which we can know to be true, but which Turing's machine cannot.

- The mechanist exposes a different thesis claiming, for example, that the human mind is a Turing machine with formal specification $S'$.

- Lucas produces the Gödel sentence for $S'$, and so on, until, presumably, the mechanist does not give up.

Lucas also deals with the concept of completability. That is, why can't we simply add the Gödel sentence to the list of theorems that $S$-machines can produce? Doing so presumably will give those machines what they lack compared to human minds. The answer is that even if we add the Gödel sentence to the $S$ machines

Lucas can simply produce a new Gödel sentence for these updated machines $S'$, i.e. a sentence that we can see is true but the new machines cannot, and so on ad infinitum. In short, as Lucas [1990] states in paragraph 9: "It is very natural ... to respond by including the Gödelian sentence in the machine, but of course that makes a different machine with a different Gödel sentence all of its own ".

Lucas [1990] observed, "although some degree of idealization seems allowable in considering a mind untrammeled by mortality..., doubts remain about how far into the infinite it is permissible to stray."

In fact, a mechanist could also try "adding a Gödelizing operator, which gives, in effect a whole denumerable infinity of Gödelian sentences". That is, some may try to give a machine a method to construct an infinite number of Gödel's sentences; if this can be done, then perhaps any Gödel sentence can be produced by the machine. Lucas [1990] argues that this is not the case; a machine with such an operator will have its own Gödel sentence, one that is not on the initial list produced by the operator. This may be possible with the transition in the *transfinite.*

Although Lucas's argument is easily attackable, it has been very successful because of the large consequences it could have in the world of Mechanism and Artificial Intelligence.

Shapiro [2016] puts this debate in a more formal way and I try to sum it up here. Let $K$ be the collection of sentences in the language of first order arithmetic that can be proved by a human in general. With *proved* here, we do not mean *to be deduced in a particular formal system* since the affirmation of Lucas (and also Penrose) goes beyond any given formal system. *Proved* here means something similar to *known with unmistakable mathematical certainty, through full mathematical rigor.* We therefore call $K$ the collection of known arithmetic sentences. For convenience, let's identify the sentences with their Gödel numbers, and then think of $K$ both as a collection of sentences and a collection of natural numbers. The protagonists of the debate assume that $K$ has sharp boundaries like any other series of natural numbers and we can then investigate its arithmetic and computational properties. The mechanist therefore claim that there is a Turing machine that enumerates $K$. In other words, they state that *$K$ is enumerable in a recursive way.* Lucas and Penrose affirm that they can refute this thesis, citing the incompleteness theorems in the crucial points.

### 4.1.1 Arguments against Lucas

As Beccuti [2018] explains, Lucas's argument is based on 3 *assailable* idealizations:

1. The consistency of the human mind
2. The knowability of the consistency of the human mind
3. The knowability of the machine that should represent the human mind

For a review of the most famous criticisms of Lucas's topic, one can consult Labinaz [2016]. The most pertinent objections to Lucas' argument are based on the invalidity of one or more of the aforementioned assumptions. In fact it is possible that the human mind is a machine, but it is not a consistent machine. Or it is possible that the mind is a consistent machine but does not know that it is. Finally it is possible that the human mind is a consistent machine, that knows it is consistent, but that it is not able to accurately establish its nature as a machine (i.e. that it does not know which kind of machine it is). Here are some arguments.

**Consistency** What I consider the main objection against Lucas is exposed by Franzén [2005] who criticizes how Lucas uses the incompleteness theorems in a hurried and not very formal way. He claims that Lucas's argument is not valid because it is based on the mistaken idea that "Gödel's theorem states that in a consistent system which is strong enough to produce simple arithmetic there are formulas which cannot be proved in the system, but which *we can see to be true.*". Franzén emphasizes that the theorem does not state anything of the kind. In general, we simply have no idea whether the Gödel sentence of a system is true or not, even in cases where it is actually true. What we do know is that the Gödel sentence is true if and only if the system is consistent, and this is not provable in the system itself. When we know that the system is consistent, we also know that Gödel's sentence is true, but in general we do not know if every formal system is consistent or not. If the human mind had the ability to determine the consistency of any consistent formal system, this would certainly mean that the human mind surpasses any computer, but there is no reason to believe that this is the case.

**Can be a machine an adequate model of the mind?** Since we cannot conclude from the incompleteness theorem that the human mind surpasses any computer as far as arithmetic is concerned, we could try to draw the weaker conclusion that no machine will be an adequate model of mind in the sense that no machine can never be exactly equivalent to the human mind with regard to arithmetic ability. But this also does not follow from the incompleteness theorem. Suppose there is a "human arithmetic skill" and we continue to assume that a particular formal system $S$ embodies exactly that ability. If we know that $S$ is consistent, we will actually have a conflict with the incompleteness theorem. But again what is missing is an argument for which we should know that $S$ is consistent. Lucas introduces here the reflection: "The best we can say is that $S$ is consistent if we are". This is irrelevant because even though we know we are consistent, there is no reason why we should conclude that $S$ is consistent, unless we already know that $S$ encodes the human arithmetic skill - and why should we know? Gödel himself commented that nothing excludes the existence of a formal system $S$ that exactly encodes human arithmetic ability, although we could not recognize the axioms of $S$ as evidently true. So we further weaken the conclusion. What follows from the incompleteness theorem is that we cannot really specify any formal system $S$ such that $S$ incorporates those and only those arithmetic truths that we can know to be true. Using the second theorem, since every $S$ system for which we know that all its arithmetic theorems are true, we can produce an arithmetic assertion - an arithmetization of "S is consistent" - that we also know to be true and that it is not a theorem in $S$. We cannot specify any formal system that exhausts all our arithmetic knowledge.

**Our Gödel sentence** Benacerraf [1967] makes a well known critique of Lucas' argument. He claims that to construct the Gödel sentence for any given formal system one must have a solid understanding of the algorithm behind the system. Furthermore, the formal system that the human mind could implement is probably extremely complex, so complex that we could never get the insight into its necessary character to build our own Gödel sentence. In other words, the fact that we understand the truth of the Gödel sentence on some systems does not imply that we can construct and see the truth of our *own* Gödel sentence. If we cannot, then perhaps we are not at all different from machines; we could be very complicated Turing machines, but still Turing machines. To rephrase this objection, let us suppose that a mechanist produces a complex formal system $S$ and claims that human minds are actually $S$. Obviously, Lucas will try to produce the Gödel sentence for $S$ to show that we are not $S$. But $S$ is extremely complicated, so complicated that Lucas cannot produce Gödel sentence for $S$, and therefore cannot deny this particular mechanistic thesis. In short, according to Benacerraf, the most we can deduce from Lucas's argument is a disjunction: "either no (formal system) encodes all human arithmetical capacity – the Lucas-Penrose thought – or any system which does has no axiomatic specification which human beings can comprehend" [Wright, 1995]. One answer that Lucas made in 1996 is that, even if he fails to create the sentence for $S$, he could be helped by other mathematicians or by computers. In short, at least according to Lucas, it could be difficult, but it seems that we could, at least in principle, determine the Gödel sentence for any given system. Very questionable objection.

This argument, that the mind can be a machine but that it is not humanly possible to know which machine it is, constitutes perhaps the strongest argument against the anti-mechanist based on the incompleteness theorems.

**The Whiteley Sentence** Whiteley [1962] responded to Lucas by claiming that humans have similar limitations to what Lucas' argument attributes to machines; if so, then perhaps we are no different from machines after all. Consider, for example, the *Whiteley sentence*, i.e. "Lucas cannot consistently assert this formula". If this sentence is true, then the statement of the sentence makes Lucas inconsistent. Thus, either Lucas is inconsistent or cannot pronounce the sentence on lack of inconsistency, in which case the sentence is true and therefore Lucas is incomplete. Even Hofstadter [1975] argues against Lucas in this direction.

## 4.2 Penrose

Penrose formulated and defended Lucas' argument in two books, The Emperor's New Mind in 1989 and Shadows of the Mind in 1994. Since the latter is at least partly an attempt to improve the first, the discussion can only concentrate on the second. Penrose [1994] consists of two main parts: (a) a Gödelian argument to prove that the minds of humans are not calculable and (b) an attempt to deduce a number of statements involving consciousness and physics from (a). I will avoid treating part (b) as it is about a domain of quantum physics that goes beyond the interest of this work.

Penrose considers its version "as the central (new) core argument against the computational modelling of mathematical understanding".

Here is a summary of the new topic as explained by Chalmers [1996]:

1. suppose that "my powers of reasoning are captured by a certain formal system $F$," and, given this assumption, "consider the class of statements I may know to be true"

2. Since I know I am sound, $F$ is sound, and so is $F'$, which is simply $F$ plus the assumption (made in (1)) that I am $F$.

3. But then "I know that $G(F')$ is true, where $G(F')$ is the Gödel sentence of system $F'$.

4. However, Gödel's first incompleteness theorem shows that $F$' could not see that the Gödel sentence is true.

5. Furthermore, we can deduce that "I am $F'$" (since $F'$ is simply $F$ plus the hypothesis made in (1) that I am $F$), and we can also infer that I can see the truth of the Gödel sentence (and therefore since we are $F'$, $F'$ can see the truth of the Gödel sentence).

6. That is, we have reached a contradiction ($F'$ can both see the truth of the Gödel sentence and not see the truth of the Gödel sentence).

7. Therefore, our initial hypothesis must be false, that is, F or any other formal system cannot capture my powers of reasoning.

### 4.2.1 Arguments against Penrose

Chalmers [1996] believes that the "greatest vulnerability" with this version of the argument is step (2); in particular, he thinks the statement that we know we are sound is problematic. McCullough [1995] states that for the success of Penrose's argument, two statements must be true: (I) "Human mathematical reasoning is sound. That is, every statement that a competent human mathematician considers to be 'unassailably true' actually is true", and (II) "The fact that human mathematical reasoning is sound is itself considered to be 'unassailably true'". These statements seem implausible to McCullough who observes: "For people (such as me) who have a more relaxed attitude towards the possibility that their reasoning might be unsound, Penrose's argument doesn't carry as much weight". McDermott [1995] questions this aspect of Penrose's argument by looking at how mathematicians actually work. He states, "it is difficult to see how thinkers like these could even be remotely approximated by an inference system that chugs to a certifiably sound conclusion, prints it out, then turns itself off". For example, McDermott points out that in 1879 Kempe published a proof of the four-color theorem that was not denied until 1890 by Heawood; there appears to have been an 11-year period in which many competent mathematicians were unsound.

Penrose tries to overcome these difficulties by distinguishing between individual and correctable errors that mathematicians sometimes make and things that can be "unequivocally" true. Penrose [1994] states "If [a] robot is ... like a genuine mathematician, although it will still make mistakes from time to time, these mistakes will be correctable ... according to its own internal criteria of 'unassailable truth' ". In other words, while mathematicians are fallible, they are still valid because their errors can be distinguished from those that can be unequivocally true and can even be corrected (and any machine, if it is to imitate mathematical reasoning, must act in the same way). The basic idea is that mathematicians can make mistakes and still be sound, because only unassailable truths are important; these truths are the output of a sound system, and we don't have to worry about the rest of the mathematicians' output. Again, a very questionable objection.

## 4.3 Formalizing the proof

J. Stern proposes in [Stern, 2018] a formalized proof for Lucas-Penrose argument.

In the debate considered in this work and availing ourselves to the notion of absolute provability, Mechanism is equivalent to the claim that the **the set of absolutely provable sentences can be recursively enumerated**. That is, there exists an explicit system of axioms and rules that proves all the absolutely provable sentences. This allows us to express the Anti-Mechanism thesis into *precise formal claims.*

Let us call *APT* (Absolute Provability and Truth) a theory extending Peano Arithmetic as explained in Chapter 2 of Stern [2018] and reported below. We need to know that *APT* is a consistent theory and we want to show that *APT* proves that the mind is not a machine.

To this end, we use the predicate $K$ for absolute provability. Moreover, if $\Sigma$ is a recursive set of axioms of some theory $\mathcal{T}$, we let $\sigma$ be a natural representation of this set in a language extending the arithmetical language of, say, Peano Arithmetic. Let $\mathrm{Pr}_\sigma$ be a natural provability predicate of $\mathcal{T}$. With this notation, Mechanist thesis becomes:

$$\exists\sigma\forall x\,(Kx \leftrightarrow Pr_\sigma(x)) \tag{MEC}$$

A refutation of Mechanism would reject this claim. That is, Anti-Mechanism would be the thesis that there is no recursive set of sentences $\Sigma$ from which all absolutely provable sentences follow:

$$\neg\exists\sigma\forall x\,(Kx \leftrightarrow Pr_\sigma(x)) \tag{ANTIMEC}$$

Following the outlines of the traditional arguments by Lucas and Penrose, the refutation of Mechanism will proceed via a reductio strategy: we will assume that the absolutely provable sentences coincide with the theorems of some recursively axiomatizable theory $\mathcal{T}$ and we will derive a contradiction starting from this assumption.

That is, we will assume

$$\forall x\,(Kx \leftrightarrow Pr_\sigma(x))$$

for some $\sigma$. However, it is important to notice that throughout the reductio proof we may not assume that the human mind knows which recursively axiomatizable theory it is—this would not amount to a refutation of Mechanism but of a much stronger claim. In particular, even though the reductio argument will be carried out in APT and, implicitly, we may assume that in our reductio assumption $Pr_\sigma$ stands for the provability predicate of APT, we may not infer $Pr_\sigma(\ulcorner\phi\urcorner)$ whenever we have proved $\phi$.

Before we give the argument, we introduce some concepts. The two constitutive principles of absolute provability, formalized by a sentential predicate $K$,

$$K\ulcorner\phi\urcorner \to \phi \tag{T}$$

$$\text{if } \phi \text{ is a theorem, then so is } K\ulcorner\phi\urcorner \tag{Nec}$$

for all sentences $\phi$ of the language, are jointly inconsistent. The idea is that (T) and (Nec) implicitly assume a naive truth predicate, which in the presence of self-referential sentences leads to paradox, as Gödel and Tarski have taught us. As a consequence, the naive truth predicate has to be replaced by a non-naive truth predicate for which $\phi$ and $T\ulcorner\phi\urcorner$ are no longer equivalent or intersubstitutable in all contexts. This means making the truth predicate explicit in formulating the principles of absolute provability. We are led to the following alternative principles of absolute provability, which seem to represent the same intuitive notion as (T) and (Nec):

$$\forall x(Kx \to Tx) \tag{$T_K$}$$

$$\text{if } T\ulcorner\phi\urcorner \text{ is a theorem, then so is } K\ulcorner\phi\urcorner \tag{T-Nec}$$

Whether any paradox will arise now only depends on the theory of truth we adopt.

We need other two definitions. First, we need the truth predicate to distribute over a disjunction:

$$T\ulcorner\phi \vee \psi\urcorner \to T\ulcorner\phi\urcorner \vee T\ulcorner\psi\urcorner \qquad \text{for all } \phi, \psi. \tag{$\vee D$}$$

$$T\ulcorner\phi\urcorner \to \phi \qquad \text{for all } \phi. \tag{T-out}$$

APT is indeed the theory extending Peano Arithmetic by the principles (∨D) and

(T -Out), together with ($T_K$).

We can now formulate in these terms the standard Liar Sentence (LS), that is, a sentence $\lambda$ such that APT, or any other arithmetical theory extending Q in an arithmetical language containing the truth predicate $T$, proves:

$$\neg T \ulcorner \lambda \urcorner \leftrightarrow \lambda \tag{LS}$$

We can now give the argument to the effect that the mind is not a machine, reasoning in APT:

**Proof** *Assume for reductio that the mind is a machine.*

$$\forall x \, (K(x) \leftrightarrow Pr_\sigma(x))$$

By the principle ($T_K$) the reductio assumption implies

$$\forall y \, (Pr_\sigma(y) \rightarrow T(y)) \tag{1}$$

and by universal instantiation

$$Pr_\sigma \left( \ulcorner \lambda \vee \neg \lambda \urcorner \right) \rightarrow T \left( \ulcorner \lambda \vee \neg \lambda \urcorner \right) \tag{2}$$

Since $\lambda \vee \neg \lambda$ is a classical tautology it is provable independently of which axioms are assumed. Therefore $\lambda \vee \neg \lambda$ is provable relative to any set of axioms and, in particular, it is provable relative to the set of axioms at stake:

$$Pr_\sigma \left( \ulcorner \lambda \vee \neg \lambda \urcorner \right). \tag{3}$$

By (2) this yields:

$$T \ulcorner \lambda \vee \neg \lambda \urcorner \tag{4}$$

Due to ($\vee D$) the truth predicate commutes with disjunction and hence

$$T \ulcorner \lambda \urcorner \vee T \ulcorner \neg \lambda \urcorner \tag{5}$$

Because of (LS) the left disjunct of (5) is equivalent to $\neg \lambda$. But due to (T -Out) the right disjunct also implies $\neg \lambda$. We can infer

$$\neg \lambda \tag{6}$$

By (L) the latter implies $T\ulcorner\lambda\urcorner$ and thus by (T-Out) we derive

$$\lambda \tag{7}$$

This ends the reductio proof since it contradicts (6). We conclude

$$\neg\forall x\,(K(x) \leftrightarrow Pr_{\sigma}(x))$$

Moreover, we have not introduced any assumption concerning $\sigma$ and therefore we can introduce the universal quantifier

$$\forall\sigma\neg\forall x\,(K(x) \leftrightarrow Pr_{\sigma}(x)) \tag{Conclusion}$$

The latter is clearly equivalent to ANTIMEC: the mind is not a machine. □

# Chapter 5

# Second Gödel's disjunct

The second disjunct says that:

> there exist absolutely unsolvable diophantine problems.

In the same Gibbs lecture, Gödel reformulate this disjunct saying that "the mathematical world is independent of human reason, insofar as there are mathematical truths that lie outside the scope of human reason". The two formulations are equivalent since the existence of mathematical absolutely unsolvable problems implies that there are some mathematical truths that the human mind can not prove and that therefore stand outside the human reason.

Unlike the first Disjunct, there were no strong supporters who tried to prove this Disjunct, probably because of the philosophical/mathematical complexity of its definition. Moreover, the underlying hypotheses on the concepts involved (absolute provability and knowability of the idealized human) are rarely well articulated and, consequently, it is difficult to evaluate the effectiveness of the arguments. In this chapter I will not analyze the arguments in favor or against this Disjunct (like in the previous chapter) but instead dwell on the concepts of relative provability, absolute provability and truth with the aim of seeing the disjunction in terms of *provability.*

Koellner [2016b] focused on these concepts, trying to analyze the set relation between them. I summarize here his thoughts.

Let $F$ be an arbitrary formal system with the characteristic that every sentence of $F$ is true and the rules of $F$ preserve the truth; let $K$ be the set of all the sentences that are absolutely provable; and let $T$ be the set of true sentences. It should be noted that $K$ and $T$ are fixed, while $F$ is a variable term that is used to capture the notion of relative provability, being relative to the considered system. We will limit $F$ such that $F \subseteq K$ and assume that $K$ has the basic feature that $K \subseteq T$. So our initial hypotheses guarantee that

$$F \subseteq K \subseteq T.$$

The question is: can we draw more substantive conclusions regarding the relationship between $F$, $K$ and $T$? For example, are there any proper inclusion?

## 5.1 Relative Provability and Truth

The first substantive conclusion is provided by the incompleteness theorems, and concerns the relation between $F$ and $T$. The incompleteness theorems tell us that for any system $F$ there are true statements that are outside the scope of $F$. In the words of Gödel:

> No well-defined system [F] of correct axioms can comprise all [of] objective mathematics [T], since the proposition which states the consistency of the system is true, but not demonstrable in the system. [Gödel, 1951]

Hence, for any $F$ we have:

$$F \subsetneq T$$

This is a clear and definite statement because the concepts involved (relative provability and arithmetic truth) are clear and defined. In contrast to the notions of $F$ and $T$, the notion $K$ of absolute provability is less clear. However, we must have that for every F:

$$F \subsetneq K \quad \text{or} \quad K \subsetneq T.$$

That is, the absolute provability exceeds the relative provability (compared to any $F$), or the truth exceeds the absolute provability. But to arrive at a more precise conclusion, i.e. for example which inclusions are proper, we require a better understanding of the nature of $K$.

## 5.2 Absolute Provability

Although the incompleteness theorems show that certain statements are undecidable with respect to particular systems (and therefore $F \subsetneq T$), it is not clear if some of these statements are absolutely undecidable ($K \subsetneq T$). Gödel certainly thought that it was not derivable from his theorems:

> [These statements are] not at all absolutely undecidable; rather, one can always pass to "higher" systems in which the sentence in question is decidable. (Some sentences, of course, nevertheless remain undecidable.) In particular, for example, it turns out that analysis is a system higher in this sense than number theory, and the axiom system of set theory is higher still than analysis. [Gödel, 1986]

Here the concept of absolute provability seems to be understood as what can be proved by some set of well-justified axioms.

But instead of trying to give a substantial analysis of the concept of absolute provability, Koellner [2016b] tries to work rather with its structural properties. The concept of absolute provability will be understood here in a highly idealized sense. For example, let us assume that it satisfies the following principles (in which we have used the "$K$" symbol as an operator):

(1) $K\varphi$, where $\varphi$ is a first-order logical validity.
(2) $(K(\varphi \to \psi) \wedge K\varphi) \to K\psi$
(3) $K\varphi \to \varphi$
(4) $K\varphi \to KK\varphi$

The first of these principles - known as logical omniscience - reveals that $K$ is considered in a highly idealized sense since some of the logical validities are too long for a real agent to understand them. This phenomenon - for which we capture the arithmetical truths by raising ourselves to higher concepts - gives us some clues that perhaps absolute provability exceeds *all* the forms of relative provability. But contrary to appearances it does not establish it completely. So maybe there is a "master system", $F*$, such that the relative provability with respect to $F*$ coincides with the absolute provability. What we can conclude is simply that if such a system exists, then we will never know (in the sense of being able to absolutely prove) that all its axioms are true. This is precisely what Gödel had in mind when he wrote the following about his incompleteness theorems:

> For, it makes it impossible that someone should set up a certain well-defined system of axioms and rules and consistently make the following assertion about it: All of these axioms and rules I perceive (with mathematical certitude) to be correct, and moreover I believe that they contain all of mathematics. If someone makes such a statement he contradicts himself. For if he perceives the axioms under consideration to be correct, he also perceives (with the same certainty) that they are consistent. Hence he has a mathematical insight not derivable from his

> axioms. [Gödel, 1951]

In other words, the incompleteness theorems allow us to conclude that $F \subsetneq T$ but not to conclude that $F \subsetneq K$. But we can draw a **conditional conclusion** to the effect that if the soundness of $F$ is absolutely provable, then

$$F \subsetneq K.$$

But how do we move to a conclusion without such strong conditions? The problem is that for all we know, we have not ruled out the possibility that there is a $F$ that coincides with $K$. Gödel was aware of this possibility:

> However, as to subjective mathematics, it is not precluded that there should exist a finite rule producing all its evident axioms. However, if such a rule exists, we with our human understanding could certainly never know it to be such, that is, we could never know with mathematical certainty that all propositions it produces are correct. For this (or the consequence concerning the consistency of the axioms) would constitute a mathematical insight not derivable from the axioms [and] rules under consideration, contrary to the assumption [Gödel, 1951]

In other words it could actually happen that there is an $F$ such that $F = K$. We have only shown that if there is such a $F$ then it must be "hidden" in the sense that we cannot absolutely prove that it has this characteristic. We can only reach a conditional conclusion, namely

$$\exists F \text{ s.t. } F = K \implies K \subsetneq T$$

Which implies that there are $\phi$ in $T$ that neither $\phi$ nor $\neg\phi$ are in $K$. Gödel says:

> [I]f the human mind were equivalent to a finite machine, then objective mathematics not only would be incompletable in the sense of not being contained in any well-defined axiomatic system, but moreover there would exist absolutely unsolvable diophantine problems of the type described above, where the epithet "absolutely" means that they would be undecidable, not just within some particular axiomatic system, but by any mathematical proof that the human mind can conceive. Gödel [1951]

then reformulating in a disjunctive form:

$$\forall F, F \subseteq K \text{ or } K \subseteq T$$

Gödel then reaches the famous formulation of the disjunction (Section 3.1.1).

In the terms of this section, either absolute provability overcomes all forms of relative provability or there are absolutely undecidable arithmetic statements.

## 5.3 Comparison with machines and human minds

So far we have talked about the concepts of relative provability, absolute provability and truth by setting aside the concepts of idealized finite machine and idealized human mind. But these latter concepts appear in Gödel's quotations and are central to the standard formulations of the disjunction. We can notice that the concept of an idealized finite machine closely corresponds to the concept of relative provability, and the concept of an idealized human mind closely corresponds to the concept of absolute provability.

The first correspondence is well established and easy to explain. Speaking of relative provability we mean the provability with respect to a recursively axiomatizable formal system, which is a precise mathematical concept. As explained at the beginning of Chapter 2, this is part of a group of precise and formal concepts that have all been shown to be equivalent to the concept of being computable by an effective procedure. Moreover, this concept had the advantage that it seemed to provide an adequate conceptual analysis of the informal concept of computability. So we can accept this analysis and take the concept of ”calculable from an idealized finite machine” co-extended with the concept of ”relative provability”.

The second correspondence is more difficult because neither the concept of idealized human mind nor the concept of absolute provability are precise. In fact, I think both concepts are problematic (like Koellner [2014] emphasizes several times). What is important for our purposes is that the idealizing hypotheses made on the concept of an idealized human mind are parallel to those made on the concept of absolute provability. And so, for our purposes, it is safe to assume (following Gödel) that these two concepts are co-extended.

## 5.4 Gödel's opinion

We can summarize this discussion as follows: Gödel thought he was able to establish the disjunctive conclusion, but he did not think he was able to establish the first disjunct (in the sense that for every F, $F \subseteq K$) because he was aware of not being able to exclude the possibility that a master system $F*$ actually existed such that $F* = K$; he could only show that *if* such a master system $F*$ existed, then it must be ”hidden” in the sense that we would never be able to absolutely show that it produced only truth.

# Chapter 6

# Conclusion

In this work I presented the significant steps of the debate on the mechanism of the human mind based on Gödel's theorems. Starting from these theorems, I have analyzed how one can deduct a disjunction for which either the capabilities of the human mind infinitely exceed those of the machines, or else there exist absolutely unsolvable problems. In this conclusion, I would like to briefly analyze the relationship between the Disjunction and artificial intelligence, which represents a main theme of the current technological world and that is also the reason why I became passionate about this topic.

In the context of the **mechanism** of the human mind, the Disjunction tells us that on the one hand the human mind surpasses every finite machine and we cannot therefore represent its capabilities with algorithms and with a computer (*Anti-Mechanism*); on the other hand, it tells us that we can *mechanize* the mind but there are truths that are absolutely unknowable to us, including our own nature. Given this, let us analyze how can we use this Disjunction in the context of artificial intelligence (AI).

In Chapter 3, I analyzed the hypotheses underlying the debate and how the various elements are treated from a philosophical point of view. In particular, it is important to remember that when we talk about computer we mean *Turing machines* and when we speak about the human mind we mean an *idealized human mind*. Under these assumptions, the Disjunction is provable and accepted. But what happens when we try to consider the real world with real computers and we want to focus on AI? As we saw in Chapter 2, we can distinguish two different types of AI: on the one hand there is **Strong AI** (or General AI) that seeks to create artificial humans, i.e. machines that have all the mental powers we have, including phenomenal consciousness, and on the other hand **Weak AI** that seeks to build information-processing machines that only *appear* to have the full mental repertoire of human persons.

In the context of the **Disjunction**, since we talk about Mechanism it is clear that the discussion is limited only to Strong AI because both theories speak of being able to represent *all* human cognitive abilities through effective procedures (see Chapter 2). We can then move on to the question that most affects the community around AI:

$$\text{Is it possible to } \textit{build} \text{ a Strong Artificial Intelligence?} \tag{D}$$

To answer (D), we can start by analyzing the Disjunction: imagine that we want to follow one or the other Disjunct.

If we consider the first Disjunct, as we did in Chapter 4, then the answer is trivial: there exist capabilities of the human mind that cannot be encapsulated in a Turing machine and, consequently, in a computer. The answer to (D) would therefore be ”no”. If we consider the second Disjunct, we would accept that absolutely unknowable truths exist. For the second incompleteness theorem, one of these unknowable truths is the consistency of man itself. As Benacerraf said, ”if I am a Turing Machine then I am barred by my very nature from obeying Socrates’ profound philosophical injunction: Know thyself”. We could therefore infer that we could encapsulate all of our capabilities in a Turing machine but we could not know its truths and its original nature. How could we consequently build it if we do not know its *true* nature?

Hence, I can summarize my thoughts about the Disjunction with respect to the possibility of realizing a Strong AI as:

- it may be that we are something more than computing machines and **it cannot exists** one machine that has all the capabilities of our mind

- or it may be that we are nothing but computing machines, and in this case we will never be able to understand exactly what type of machine we are and consequently **we cannot build** it.

Today there are many companies that invest large amounts of capital in research and development for *building* artificial intelligence. The AI market magnitude is today of the order of tens of billions of dollars. Hence, am I concluding that the investments made today in AI are useless? Absolutely not.

Most of the companies treat only the ”Weak” AI . The Disjunction only refers to Strong AI and it does not exclude the possibility of existence of Weak AI. The impossibility of creating a machine that is the ”perfect copy” of a human does not exclude the possibility to create machines that outperform humans in many (or even all) daily tasks. In fact today in many fields like computer vision, natural language processing or information retrieval, there are models that already have super-human performances in very specific tasks. These machines can help people in the most repetitive jobs, allowing them to concentrate where there is greater added value.

In addition to these topic-focus companies, there are enterprises and research centers that try to go beyond and aim to create a Strong AI, often called Artificial General Intelligence (AGI). An example is *Open AI*, in my opinion the main company that is moving in this direction today. Let us note, however, that they themselves define AGI as ”*highly autonomous systems that outperform humans at most economically valuable work*”. This is very different from the true essence of Strong AI: the attention is shifted from ”encapsulating the capacities of the human mind” to ”outperforming humans in the tasks commonly associated with their jobs”. I firmly believe that this is possible, given that this type of AI can be seen as a generalization of Weak AI rather than AGI.

In conclusion, we have seen how the Gödel’s Incompleteness Theorems led to a disjunction that is becoming a landmark in the debate on the mechanism of the human mind. From this, I have inferred that Strong AI is impossible to reach but instead we can achieve the development of a generalized Weak AI.

# Bibliography

F. Beccuti. La disgiunzione di gödel. In *N18 / APhEx*. Aphex, 2018.

P. Benacerraf. God, the devil and gödel. *The Monist, LI, pp. 9-32*, 1967.

G. Boolos. Journal of symbolic logic 47. *LPS –Logic and Philosophy of Science*, page 191–196, 1982.

S. Bringsjord and N. S. Govindarajulu. Artificial intelligence. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition, 2018.

D. J. Chalmers. *Minds, Machines and Mathematics*. Psyche, 1996.

A. Church. An unsolvable problem of elementary number theory. *American Journal of Mathematics*, 58:345–363, 1936.

V. Fano and P Graziani. Gödel and the fundamental incompleteness of human self-knowledge. *LPS –Logic and Philosophy of Science*, IX(1):263–274, 2011.

S. Feferman. Are there absolutely unsolvable problems? gödel's dichotomy. *Philosophia Mathematica XIV, 2, pp. 134–152*, 2006.

S. Feferman and R. Solovay. Introductory note to 1972a. In *S. Feferman et al. (eds.) Kurt Gödel. Collected Works. Volume II: Publications 1938–1974*, pages 281–304. Oxford: Oxford University Press, 1990.

T. Franzén. Gödel's theorem: An incomplete guide to its use and abuse. *Wellesley: A.K. Peters.*, 2005.

K. Gödel. Some basic theorems on the foundations of mathematics and their implications. In *1990 Collected Works, Volume II, pp. 304-323*. New York, Oxford University Press, 1951.

K. Gödel. *In Solomon Feferman, ed., Kurt Gödel, Collected Works, Volume I Publications 1929–1936*. Oxford University Press, New York, 1986.

J. Haugeland. *Artificial Intelligence: The Very Idea.* Cambridge: MIT Press, 1985.

D. Hofstadter. *Gödel, Escher Bach.* New York, Basic Books, 1975.

L. Horsten and P. Welch. *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledgey.* Hoxford University Press, 2016.

P. Koellner. On the Question of Absolute Undecidability. *Philosophia Mathematica*, 14(2):153–188, 2006.

P. Koellner. On the question of whether the mind can be mechanized. *Forthcoming, Proceedings of the London Mathematical Society*, 2014.

P. Koellner. The continuum hypothesis. *Zalta E.N.*, 2016a.

P. Koellner. Gödel's disjunction. In L. Horsten and P. Welch, editors, *Gödel's Disjunction*, chapter 8. Hoxford University Press, 2016b.

P. Labinaz. Introduzione. In *Labinaz P. (a cura di), J. R. Lucas Against Mechanism.* Milano, Mimesis, 2016.

J. R. Lucas. Minds, machines and gödel. *Philosophy, XXXVI, pp. 112-137*, 1961.

J. R. Lucas. Mind, machines and gödel: A retrospect. *A paper read to the Turing Conference at Brighton on April 6 th*, 1990.

G. Luger and W. Stubblefield. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving.* Redwood, CA: Benjamin Cummings, 1993.

D. McCullough. Can humans escape godel? a review of shadows of the mind by roger penrose. *PSYCHE: An Interdisciplinary Journal of Research On Consciousness*, 2:57–65, 1995.

D. McDermott. [star] penrose is wrong. *PSYCHE: An Interdisciplinary Journal of Research On Consciousness*, 2:66–82, 1995.

R. Murawski. *Recursive Functions and Metamathematics: Problems of Completeness and Decidability, Gödel's Theorems.* Dordrecht: Kluwer, 1999.

R. Penrose. *The Emperor's New Mind.* Oxford, Oxford University Press, 1989.

R. Penrose. *Shadows of the Mind.* Oxford, Oxford University Press, 1994.

P. Raatikainen. On the philosophical relevance of gödel's incompleteness theorems. *Revue Internationale de Philosophie, 59: 513–534*, 2005.

P. Raatikainen. Gödel's disjunction: The scope and limits of mathematical knowledge. *History and Philosophy of Logic*, 2018a.

P. Raatikainen. Gödel's incompleteness theorems. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition, 2018b.

S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach (2nd Edition)*. Prentice Hall, 2002.

J. Searle. Roger penrose, kurt gödel, and the cytoskeletons. In *Searle: Mystery of Consciousness*, page 55–93. New York: New York Review of Books, 1997.

S. Shapiro. Idealization, mechanism, and knowability. In L. Horsten and P. Welch, editors, *Gödel's Disjunction*, chapter 8. Hoxford University Press, 2016.

P. Smith. *An Introduction to Gödel's Theorems*. Cambridge: Cambridge University Press, 2007.

A. Sokal and J. Bricmont. *Fashionable Non-sense: Postmodern Intellectuals' Abuse of Science, Picador Books*. New York, 1998.

J. Stern. *Proving that the Mind Is Not a Machine?* Thought. A Journal of Philosophy 7(2): 81–90., 2018.

A. M. Turing. On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42:230–265, 1936–1937.

H. Wang. A logical journey. *Cambridge (MA), The MIT Press*, 1996.

C. H. Whiteley. Minds, machines and godel : A reply to mr lucas. *Philosophy*, 37 (139):61–62, 1962.

T. Williamson. Absolute provability and safe knowledge of axioms. In *Gödel's Disjunction,*. Oxford, Oxford University Press, 2016.

P. Winston. *Artificial Intelligence*. eading, MA: Addison-Wesley, 1992.

C. Wright. Intuitionists are not turing machines. *Philosophia Mathematica 3(1):86-102*, 1995.