Universality and uncomputability

Boaz Barak

**Disclaimer:** The MS-Word version of this text has significant formatting issues compared to the PDF and HTML versions.

# Universality and uncomputability

### 

* The universal machine/program - “one program to rule them all”
* A fundamental result in computer science and mathematics: the existence of uncomputable functions.
* The *halting problem*: the canonical example of an uncomputable function.
* Introduction to the technique of *reductions*.
* Rice’s Theorem: A “meta tool” for uncomputability results, and a starting point for much of the research on compilers, programming languages, and software verification.

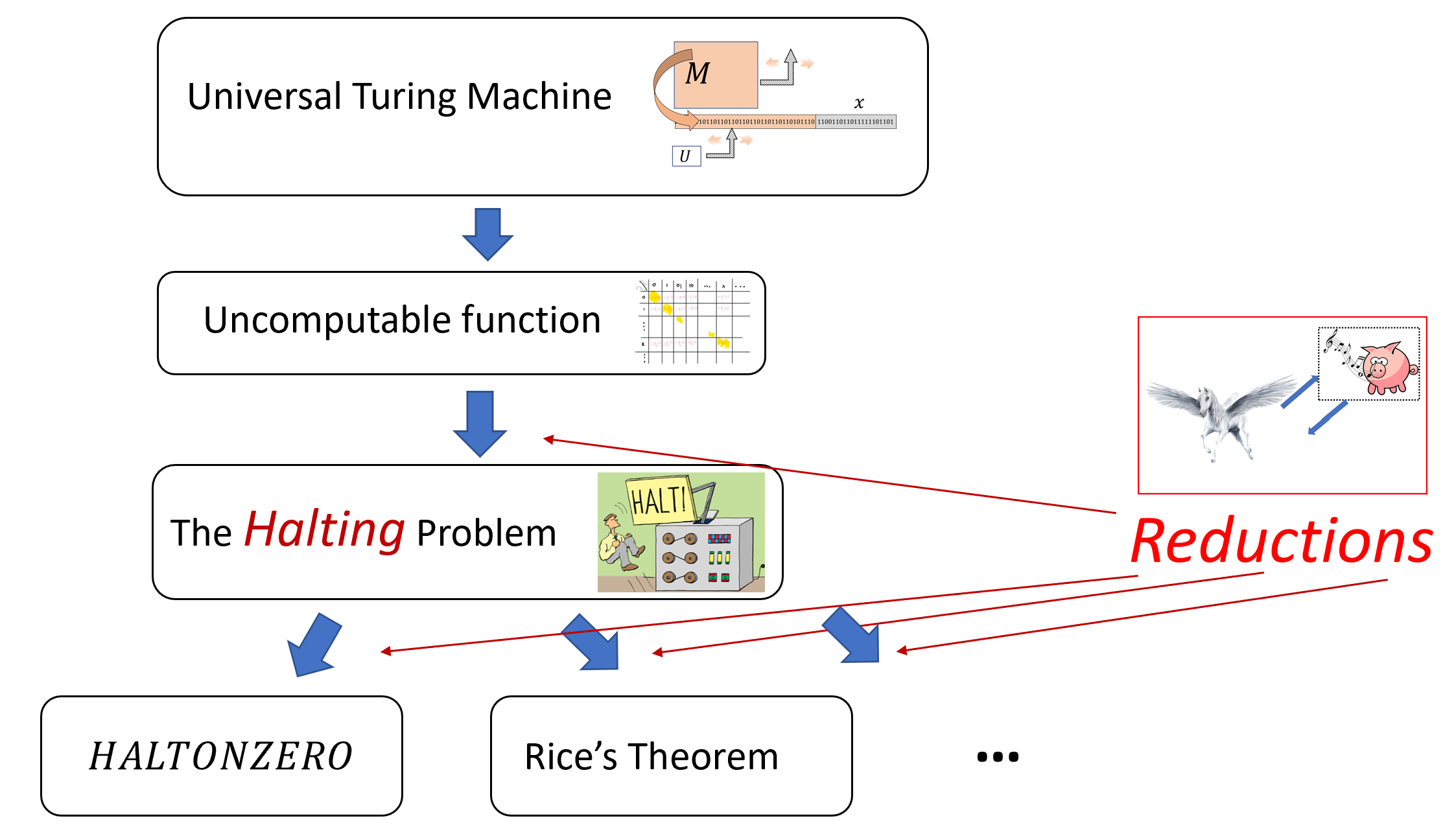
*“A function of a variable quantity is an analytic expression composed in any way whatsoever of the variable quantity and numbers or constant quantities.”*, Leonhard Euler, 1748.

*“The importance of the universal machine is clear. We do not need to have an infinity of different machines doing different jobs. … The engineering problem of producing various machines for various jobs is replaced by the office work of ‘programming’ the universal machine”*, Alan Turing, 1948

One of the most significant results we showed for Boolean circuits (or equivalently, straight-line programs) is the notion of *universality*: there is a single circuit that can evaluate all other circuits. However, this result came with a significant caveat. To evaluate a circuit of gates, the universal circuit needed to use a number of gates *larger* than . It turns out that uniform models such as Turing machines or NAND-TM programs allow us to “break out of this cycle” and obtain a truly *universal Turing machine* that can evaluate all other machines, including machines that are more complex (e.g., more states) than itself. (Similarly, there is a *Universal NAND-TM program* that can evaluate all NAND-TM programs, including programs that have more lines than .)

It is no exaggeration to say that the existence of such a universal program/machine underlies the information technology revolution that began in the latter half of the 20th century (and is still ongoing). Up to that point in history, people have produced various special-purpose calculating devices such as the abacus, the slide ruler, and machines that compute various trigonometric series. But as Turing (who was perhaps the one to see most clearly the ramifications of universality) observed, a *general purpose computer* is much more powerful. Once we build a device that can compute the single universal function, we have the ability, *via software*, to extend it to do arbitrary computations. For example, if we want to simulate a new Turing machine , we do not need to build a new physical machine, but rather can represent as a string (i.e., using *code*) and then input to the universal machine .

Beyond the practical applications, the existence of a universal algorithm also has surprising theoretical ramifications, and in particular can be used to show the existence of *uncomputable functions*, upending the intuitions of mathematicians over the centuries from Euler to Hilbert. In this chapter we will prove the existence of the universal program, and also show its implications for uncomputability, see universalchapoverviewfig



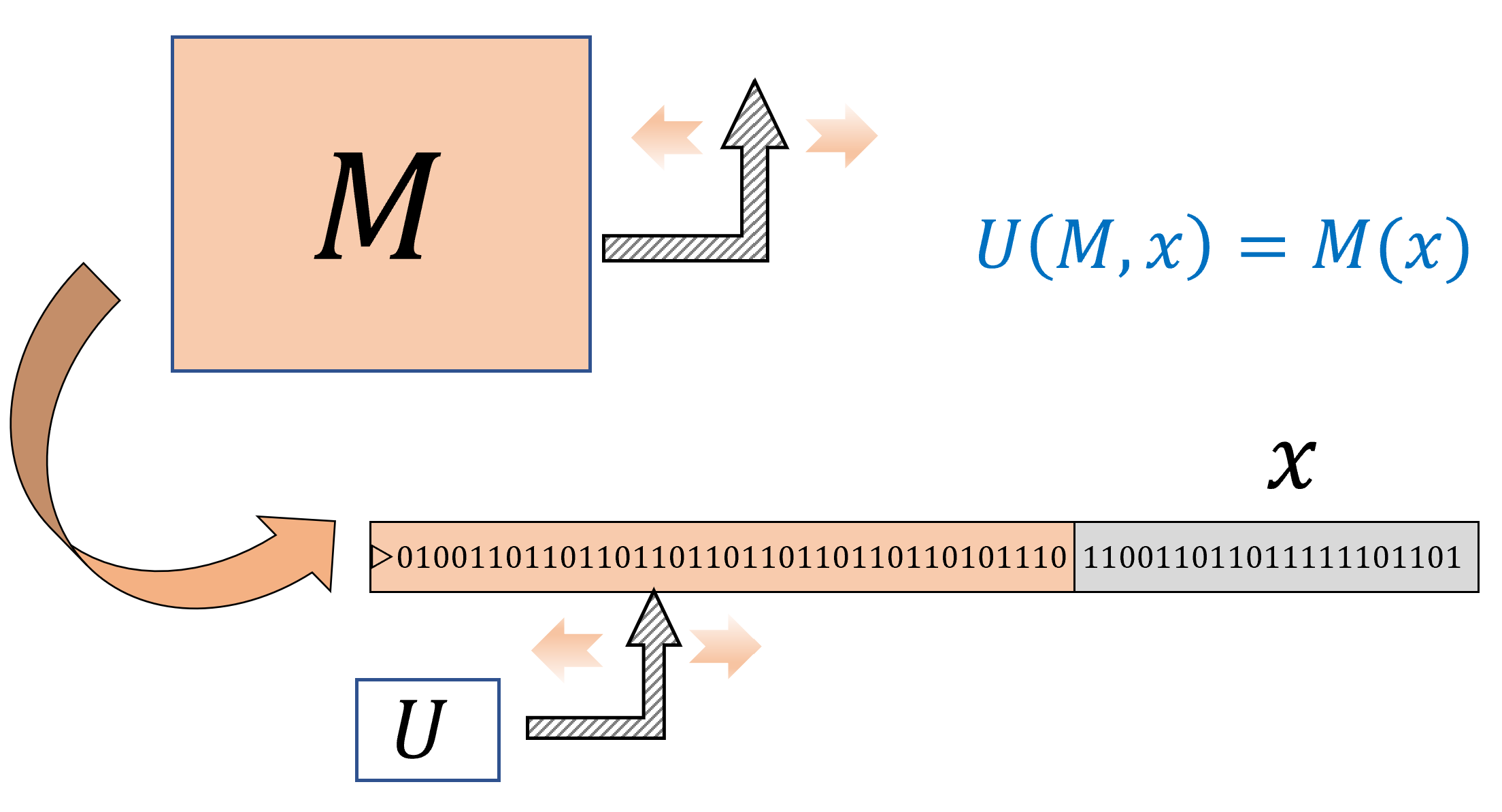
In this chapter we will show the existence of a *universal Turing machine* and then use this to derive first the existence of *some* uncomputable function. We then use this to derive the uncomputability of Turing’s famous “halting problem” (i.e., the function), from which we a host of other uncomputability results follow. We also introduce *reductions*, which allow us to use the uncomputability of a function to derive the uncomputability of a new function .

## Universality or a meta-circular evaluator

We start by proving the existence of a *universal Turing machine*. This is a single Turing machine that can evaluate *arbitrary* Turing machines on *arbitrary* inputs , including machines that can have more states and larger alphabet than itself. In particular, can even be used to evaluate itself! This notion of *self reference* will appear time and again in this course, and as we will see, leads to several counter-intuitive phenomena in computing.

There exists a Turing machine such that on every string which represents a Turing machine, and , .

That is, if the machine halts on and outputs some then , and if does not halt on (i.e., ) then .



A *Universal Turing Machine* is a single Turing Machine that can evaluate, given input the (description as a string of) arbitrary Turing machine and input , the output of on . In contrast to the universal circuit depicted in universalcircfig, the machine can be much more complex (e.g., more states or tape alphabet symbols) than .

There is a *“universal”* algorithm that can evaluate arbitrary algorithms on arbitrary inputs.

Once you understand what the theorem says, it is not that hard to prove. The desired program is an *interpreter* for Turing machines. That is, gets a representation of the machine (think of it as source code), and some input , and needs to simulate the execution of on .

Think of how you would code in your favorite programming language. First, you would need to decide on some representation scheme for . For example, you can use an array or a dictionary to encode ’s transition function. Then you would use some data structure, such as a list, to store the contents of ’s tape. Now you can simulate step by step, updating the data structure as you go along. The interpreter will continue the simulation until the machine halts.

Once you do that, translating this interpreter from your favorite programming language to a Turing machine can be done just as we have seen in chapequivalentmodels. The end result is what’s known as a “meta-circular evaluator”: an interpreter for a programming language in the same one. This is a concept that has a long history in computer science starting from the original universal Turing machine. See also lispinterpreterfig.

### Proving the existence of a universal Turing Machine

To prove (and even properly state) universaltmthm, we need to fix some representation for Turing machines as strings. For example, one potential choice for such a representation is to use the equivalence betwen Turing machines and NAND-TM programs and hence represent a Turing machine using the ASCII encoding of the source code of the corresponding NAND-TM program . However, we will use a more direct encoding.

Let be a Turing machine with states and a size alphabet (we use the convention ,, , ). We represent as the triple where is the table of values for :

where each value is a triple with , and a number encoding one of . Thus such a machine is encoded by a list of natural numbers. The *string representation* of is obtained by concatenating prefix free representation of all these integers. If a string does not represent a list of integers in the form above, then we treat it as representing the trivial Turing machine with one state that immediately halts on every input.

The details of the representation scheme of Turing machines as strings are immaterial for almost all applications. What you need to remember are the following points:

1. We can represent every Turing machine as a string.
2. Given the string representation of a Turing machine and an input , we can simulate ’s execution on the input . (This is the content of universaltmthm.)

An additional minor issue is that for convenience we make the assumption that *every* string represents *some* Turing machine. This is very easy to ensure by just mapping strings that would otherwise not represent a Turing machine into some fixed trivial machine. This assumption is not very important, but does make a few results (such as Rice’s Theorem: rice-thm) a little less cumbersome to state.

Using this representation, we can formally prove universaltmthm.

We will only sketch the proof, giving the major ideas. First, we observe that we can easily write a *Python* program that, on input a representation of a Turing machine and an input , evaluates on . Here is the code of this program for concreteness, though you can feel free to skip it if you are not familiar with (or interested in) Python:

# constants  
def EVAL(δ,x):  
 '''Evaluate TM given by transition table δ  
 on input x'''  
 Tape = ["▷"] + [a for a in x]  
 i = 0; s = 0 # i = head pos, s = state  
 while True:  
 s, Tape[i], d = δ[(s,Tape[i])]  
 if d == "H": break  
 if d == "L": i = max(i-1,0)  
 if d == "R": i += 1  
 if i>= len(Tape): Tape.append('Φ')  
  
 j = 1; Y = [] # produce output  
 while Tape[j] != 'Φ':  
 Y.append(Tape[j])  
 j += 1  
 return Y

On input a transition table this program will simulate the corresponding machine step by step, at each point maintaining the invariant that the array Tape contains the contents of ’s tape, and the variable s contains ’s current state.

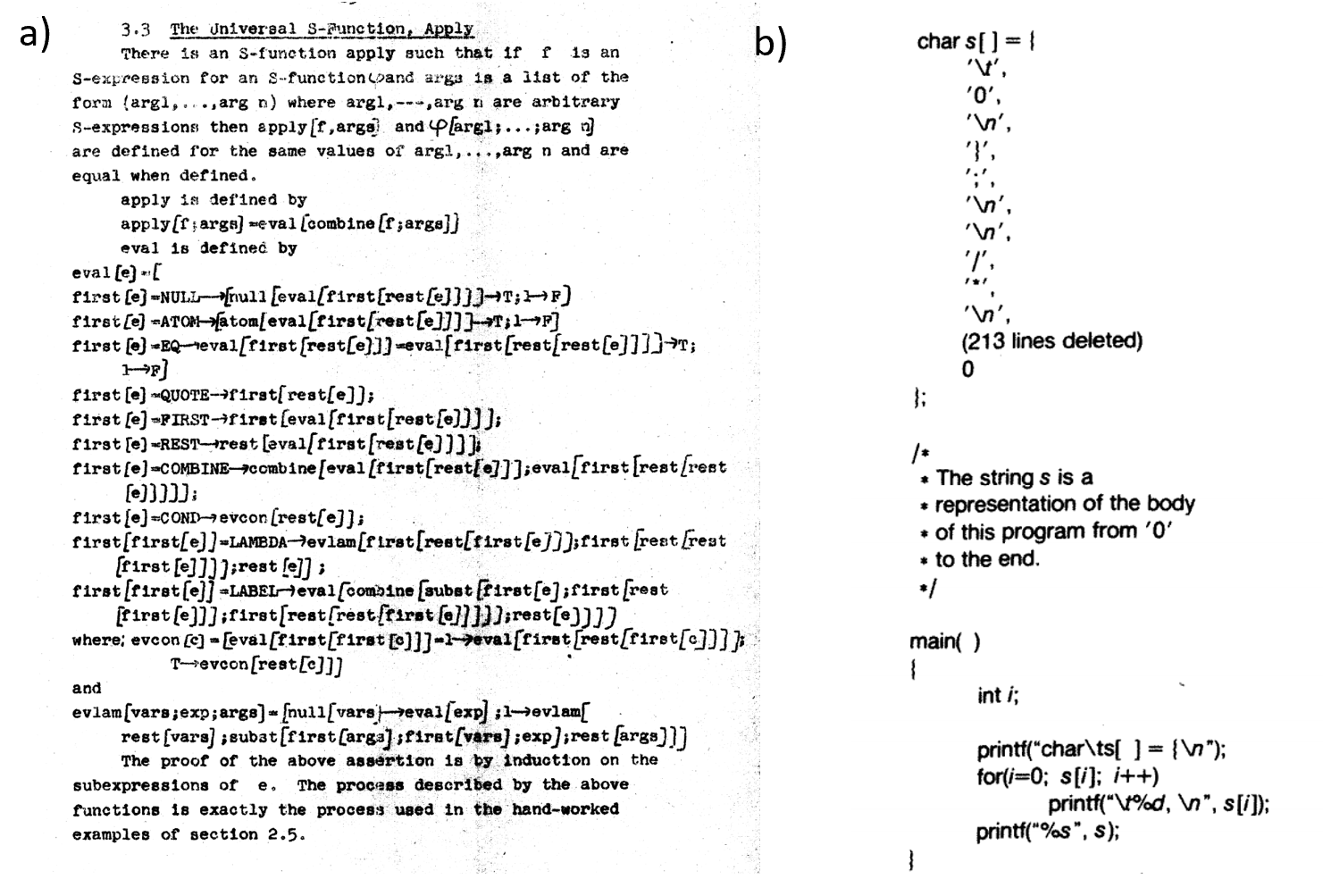
The above does not prove the theorem as stated, since we need to show a *Turing machine* that computes rather than a Python program. With enough effort, we can translate this Python code line by line to a Turing machine. However, to prove the theorem we don’t need to do this, but can use our “eat the cake and have it too” paradigm. That is, while we need to evaluate a Turing machine, in writing the code for the interpreter we are allowed to use a richer model such as NAND-RAM since it is equivalent in power to Turing machines per RAMTMequivalencethm).

Translating the above Python code to NAND-RAM is truly straightforward. The only issue is that NAND-RAM doesn’t have the *dictionary* data structure built in, which we have used above to store the transition function δ. However, we can represent a dictionary of the form as simply a list of pairs. To compute we can scan over all the pairs until we find one of the form in which case we return . Similarly we scan the list to update the dictionary with a new value, either modifying it or appending the pair at the end.

The argument in the proof of universaltmthm is a very inefficient way to implement the dictionary data structure in practice, but it suffices for the purpose of proving the theorem. Reading and writing to a dictionary of values in this implementation takes steps, but it is in fact possible to do this in steps using a *search tree* data structure or even (for “typical” instances) using a *hash table*. NAND-RAM and RAM machines correspond to the architecture of modern electronic computers, and so we can implement hash tables and search trees in NAND-RAM just as they are implemented in other programming languages.

Since universal Turing

### Implications of universality (discussion)



**a)** A particularly elegant example of a “meta-circular evaluator” comes from John McCarthy’s 1960 paper, where he defined the Lisp programming language and gave a Lisp function that evaluates an arbitrary Lisp program (see above). Lisp was not initially intended as a practical programming language and this example was merely meant as an illustration that the Lisp universal function is more elegant than the universal Turing machine. It was McCarthy’s graduate student Steve Russell who suggested that it can be implemented. As McCarthy later recalled, *“I said to him, ho, ho, you’re confusing theory with practice, this eval is intended for reading, not for computing. But he went ahead and did it. That is, he compiled the eval in my paper into IBM 704 machine code, fixing a bug, and then advertised this as a Lisp interpreter, which it certainly was”.* **b)** A self-replicating C program from the classic essay of Thompson [@thompson1984reflections].

There is more than one Turing machine that satisfies the conditions of universaltmthm, but the existence of even a single such machine is already extremely fundamental to both the theory and practice of computer science. universaltmthm’s impact reaches beyond the particular model of Turing machines. Because we can simulate every Turing Machine by a NAND-TM program and vice versa, universaltmthm immediately implies there exists a universal NAND-TM program such that for every NAND-TM program . We can also “mix and match” models. For example since we can simulate every NAND-RAM program by a Turing machine, and every Turing Machine by the calculus, universaltmthm implies that there exists a expression such that for every NAND-RAM program and input on which , if we encode as a -expression (using the -calculus encoding of strings as lists of ’s and ’s) then evaluates to an encoding of . More generally we can say that for every and in the set Turing Machines, RAM Machines, NAND-TM, NAND-RAM, -calculus, JavaScript, Python, of Turing equivalent models, there exists a program/machine in that computes the map for every program/machine .

The idea of a “universal program” is of course not limited to theory. For example compilers for programming languages are often used to compile *themselves*, as well as programs more complicated than the compiler. (An extreme example of this is Fabrice Bellard’s [Obfuscated Tiny C Compiler](https://bellard.org/otcc/) which is a C program of 2048 bytes that can compile a large subset of the C programming language, and in particular can compile itself.) This is also related to the fact that it is possible to write a program that can print its own source code, see lispinterpreterfig. There are universal Turing machines known that require a very small number of states or alphabet symbols, and in particular there is a universal Turing machine (with respect to a particular choice of representing Turing machines as strings) whose tape alphabet is and has fewer than states (see uncomputablebibnotes).

## Is every function computable?

In NAND-univ-thm, we saw that NAND-CIRC programs can compute every finite function . Therefore a natural guess is that NAND-TM programs (or equivalently, Turing Machines) could compute every infinite function . However, this turns out to be *false*. That is, there exists a function that is *uncomputable*!

The existence of uncomputable functions is quite surprising. Our intuitive notion of a “function” (and the notion most mathematicians had until the 20th century) is that a function defines some implicit or explicit way of computing the output from the input . The notion of an “uncomputable function” thus seems to be a contradiction in terms, but yet the following theorem shows that such creatures do exist:

### 

There exists a function that is not computable by any Turing machine.

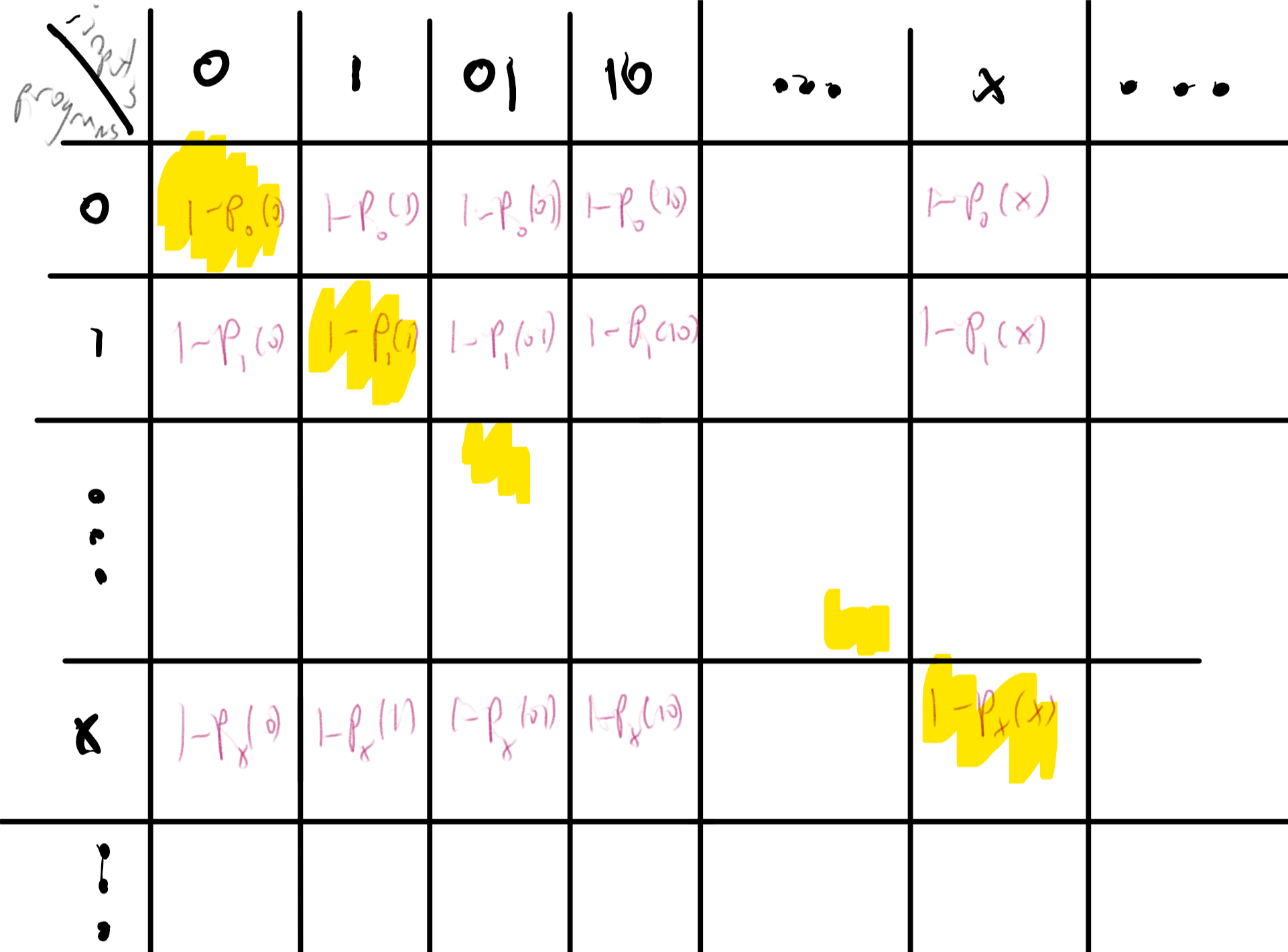
### 

The idea behind the proof follows quite closely Cantor’s proof that the reals are uncountable (cantorthm), and in fact the theorem can also be obtained fairly directly from that result (see uncountablefuncex). However, it is instructive to see the direct proof. The idea is to construct in a way that will ensure that every possible machine will in fact fail to compute . We do so by defining to equal if describes a Turing machine which satisfies and defining otherwise. By construction, if is any Turing machine and is the string describing it, then and therefore does *not* compute .

The proof is illustrated in diagonal-fig. We start by defining the following function :

For every string , if satisfies **(1)** is a valid representation of some Turing machine (per the representation scheme above) and **(2)** when the program is executed on the input it halts and produces an output, then we define as the first bit of this output. Otherwise (i.e., if is not a valid representation of a Turing machine, or the machine never halts on ) we define . We define .

We claim that there is no Turing machine that computes . Indeed, suppose, towards the sake of contradiction, there exists a machine that computes , and let be the binary string that represents the machine . On one hand, since by our assumption computes , on input the machine halts and outputs . On the other hand, by the definition of , since is the representation of the machine , , hence yielding a contradiction.



We construct an uncomputable function by defining for every two strings the value which equals if the machine described by outputs on , and otherwise. We then define to be the “diagonal” of this table, namely for every . The function is uncomputable, because if it was computable by some machine whose string description is then we would get that .

There are some functions that *can not* be computed by *any* algorithm.

### 

The proof of uncomputable-func is short but subtle. I suggest that you pause here and go back to read it again and think about it - this is a proof that is worth reading at least twice if not three or four times. It is not often the case that a few lines of mathematical reasoning establish a deeply profound fact - that there are problems we simply *cannot* solve.

The type of argument used to prove uncomputable-func is known as *diagonalization* since it can be described as defining a function based on the diagonal entries of a table as in diagonal-fig. The proof can be thought of as an infinite version of the *counting* argument we used for showing lower bound for NAND-CIRC programs in counting-lb. Namely, we show that it’s not possible to compute all functions from by Turing machines simply because there are more functions like that then there are Turing machines.

As mentioned in decidablelanguages, many texts use the “language” terminology and so will call a set an [*undecidable*](https://goo.gl/3YvQvL) or *non recursive* language if the function such that is uncomputable.

## The Halting problem

uncomputable-func shows that there is *some* function that cannot be computed. But is this function the equivalent of the “tree that falls in the forest with no one hearing it”? That is, perhaps it is a function that no one actually *wants* to compute. It turns out that there are natural uncomputable functions:

### 

Let be the function such that for every string , if Turing machine halts on the input and otherwise. Then is not computable.

Before turning to prove halt-thm, we note that is a very natural function to want to compute. For example, one can think of as a special case of the task of managing an “App store”. That is, given the code of some application, the gatekeeper for the store needs to decide if this code is safe enough to allow in the store or not. At a minimum, it seems that we should verify that the code would not go into an infinite loop.

One way to think about this proof is as follows:

That is, we will use the universal Turing machine that computes to derive the uncomputability of from the uncomputability of shown in uncomputable-func. Specifically, the proof will be by contradiction. That is, we will assume towards a contradiction that is computable, and use that assumption, together with the universal Turing machine of universaltmthm, to derive that is computable, which will contradict uncomputable-func.

If a function is uncomputable we can show that another function is uncomputable by giving a way to *reduce* the task of computing to computing .

The proof will use the previously established result uncomputable-func. Recall that uncomputable-func shows that the following function is uncomputable:

where denotes the output of the Turing machine described by the string on the input (with the usual convention that if this computation does not halt).

We will show that the uncomputability of implies the uncomputability of . Specifically, we will assume, towards a contradiction, that there exists a Turing machine that can compute the function, and use that to obtain a Turing machine that computes the function . (This is known as a proof by *reduction*, since we reduce the task of computing to the task of computing . By the contrapositive, this means the uncomputability of implies the uncomputability of .)

Indeed, suppose that is a Turing machine that computes . halttof describes a Turing Machine that computes . (We use “high level” description of Turing machines, appealing to the “have your cake and eat it too” paradigm, see eatandhavecake.)

INPUT: $x\in \{0,1\}^\*$  
OUTPUT: $F^\*(x)$  
# Assume T.M. $M\_{HALT}$ computes $HALT$  
  
Let $z \leftarrow M\_{HALT}(x,x)$. # Assume $z=HALT(x,x)$.  
If{$z=0$}  
return $0$  
endif  
Let $y \leftarrow U(x,x)$ # $U$ universal TM, i.e., $y=x(x)$  
If{$y=0$}  
return $1$  
endif  
Return $0$

We claim that halttof computes the function . Indeed, suppose that (and hence ). In this case, and hence, under our assumption that , the value will equal , and hence halttof will set , and output the correct value .

Suppose otherwise that (and hence ). In this case there are two possibilities:

* **Case 1:** The machine described by does not halt on the input . In this case, . Since we assume that computes it means that on input , the machine must halt and output the value . This means that halttof will set and output .
* **Case 2:** The machine described by halts on the input and outputs some . In this case, since , under our assumptions, halttof will set and so output .

We see that in all cases, , which contradicts the fact that is uncomputable. Hence we reach a contradiction to our original assumption that computes .

### 

Once again, this is a proof that’s worth reading more than once. The uncomputability of the halting problem is one of the fundamental theorems of computer science, and is the starting point for much of the investigations we will see later. An excellent way to get a better understanding of halt-thm is to go over haltalternativesec, which presents an alternative proof of the same result.

### Is the Halting problem really hard? (discussion)

Many people’s first instinct when they see the proof of halt-thm is to not believe it. That is, most people do believe the mathematical statement, but intuitively it doesn’t seem that the Halting problem is really that hard. After all, being uncomputable only means that cannot be computed by a Turing machine.

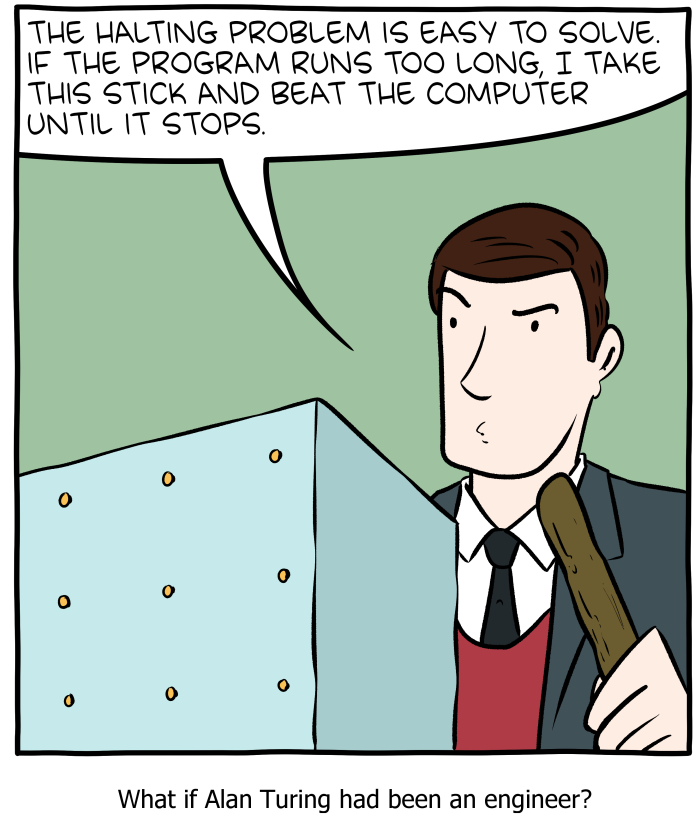
But programmers seem to solve all the time by informally or formally arguing that their programs halt. It’s true that their programs are written in C or Python, as opposed to Turing machines, but that makes no difference: we can easily translate back and forth between this model and any other programming language.

While every programmer encounters at some point an infinite loop, is there really no way to solve the halting problem? Some people argue that *they* personally can, if they think hard enough, determine whether any concrete program that they are given will halt or not. Some have even [argued](https://goo.gl/Bm4MWK) that humans in general have the ability to do that, and hence humans have inherently superior intelligence to computers or anything else modeled by Turing machines.[[1]](#footnote-53)

The best answer we have so far is that there truly is no way to solve , whether using Macs, PCs, quantum computers, humans, or any other combination of electronic, mechanical, and biological devices. Indeed this assertion is the content of the *Church-Turing Thesis*. This of course does not mean that for *every* possible program , it is hard to decide if enters an infinite loop. Some programs don’t even have loops at all (and hence trivially halt), and there are many other far less trivial examples of programs that we can certify to never enter an infinite loop (or programs that we know for sure that *will* enter such a loop). However, there is no *general procedure* that would determine for an *arbitrary* program whether it halts or not. Moreover, there are some very simple programs for which no one knows whether they halt or not. For example, the following Python program will halt if and only if [Goldbach’s conjecture](https://goo.gl/DX63q5) is false:

def isprime(p):  
 return all(p % i for i in range(2,p-1))  
  
def Goldbach(n):  
 return any( (isprime(p) and isprime(n-p))  
 for p in range(2,n-1))  
  
n = 4  
while True:  
 if not Goldbach(n): break  
 n+= 2

Given that Goldbach’s Conjecture has been open since 1742, it is unclear that humans have any magical ability to say whether this (or other similar programs) will halt or not.



[SMBC](http://smbc-comics.com/comic/halting)’s take on solving the Halting problem.

### A direct proof of the uncomputability of (optional)

It turns out that we can combine the ideas of the proofs of uncomputable-func and halt-thm to obtain a short proof of the latter theorem, that does not appeal to the uncomputability of . This short proof appeared in print in a 1965 letter to the editor of Christopher Strachey:

To the Editor, The Computer Journal.

An Impossible Program

Sir,

A well-known piece of folk-lore among programmers holds that it is impossible to write a program which can examine any other program and tell, in every case, if it will terminate or get into a closed loop when it is run. I have never actually seen a proof of this in print, and though Alan Turing once gave me a verbal proof (in a railway carriage on the way to a Conference at the NPL in 1953), I unfortunately and promptly forgot the details. This left me with an uneasy feeling that the proof must be long or complicated, but in fact it is so short and simple that it may be of interest to casual readers. The version below uses CPL, but not in any essential way.

Suppose T[R] is a Boolean function taking a routine (or program) R with no formal or free variables as its arguments and that for all R, T[R] = True if R terminates if run and that T[R] = False if R does not terminate.

Consider the routine P defined as follows

rec routine P  
§L: if T[P] go to L  
Return §

If T[P] = True the routine P will loop, and it will only terminate if T[P] = False. In each case `T[P]`` has exactly the wrong value, and this contradiction shows that the function T cannot exist.

Yours faithfully,  
C. Strachey

Churchill College, Cambridge

Try to stop and extract the argument for proving halt-thm from the letter above.

Since CPL is not as common today, let us reproduce this proof. The idea is the following: suppose for the sake of contradiction that there exists a program T such that T(f,x) equals True iff f halts on input x. (Strachey’s letter considers the no-input variant of , but as we’ll see, this is an immaterial distinction.) Then we can construct a program P and an input x such that T(P,x) gives the wrong answer. The idea is that on input x, the program P will do the following: run T(x,x), and if the answer is True then go into an infinite loop, and otherwise halt. Now you can see that T(P,P) will give the wrong answer: if P halts when it gets its own code as input, then T(P,P) is supposed to be True, but then P(P) will go into an infinite loop. And if P does not halt, then T(P,P) is supposed to be False but then P(P) will halt. We can also code this up in Python:

def CantSolveMe(T):  
 """  
 Gets function T that claims to solve HALT.  
 Returns a pair (P,x) of code and input on which  
 T(P,x) ≠ HALT(x)  
 """  
 def fool(x):  
 if T(x,x):  
 while True: pass  
 return "I halted"  
  
 return (fool,fool)

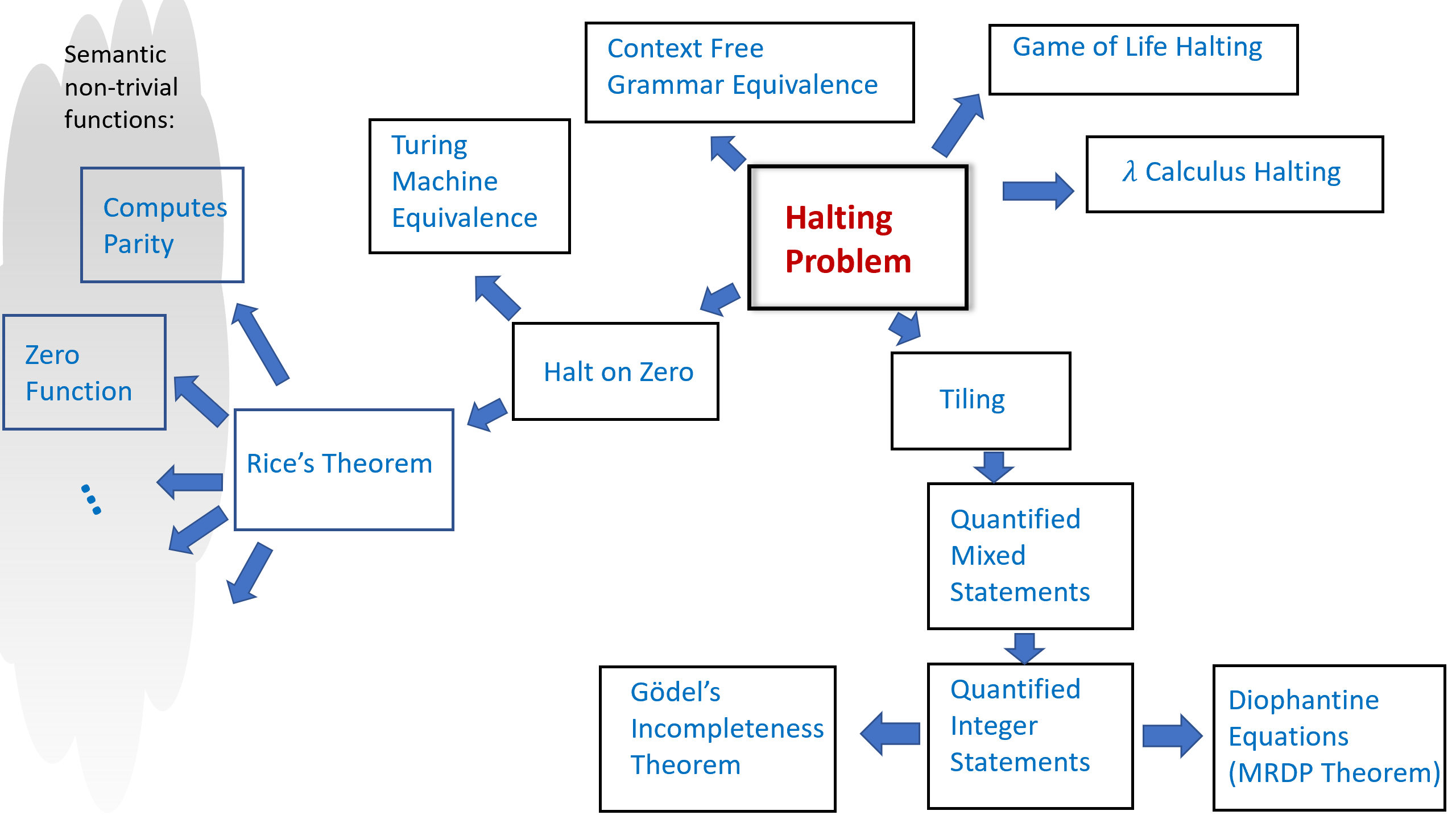
For example, consider the following Naive Python program T that guesses that a given function does not halt if its input contains while or for

def T(f,x):  
 """Crude halting tester - decides it doesn't halt if it contains a loop."""  
 import inspect  
 source = inspect.getsource(f)  
 if source.find("while"): return False  
 if source.find("for"): return False  
 return True

If we now set (f,x) = CantSolveMe(T), then T(f,x)=False but f(x) does in fact halt. This is of course not specific to this particular T: for every program T, if we run (f,x) = CantSolveMe(T) then we’ll get an input on which T gives the wrong answer to .

## Reductions

The Halting problem turns out to be a linchpin of uncomputability, in the sense that halt-thm has been used to show the uncomputability of a great many interesting functions. We will see several examples of such results in this chapter and the exercises, but there are many more such results (see haltreductions).



Some uncomputability results. An arrow from problem X to problem Y means that we use the uncomputability of X to prove the uncomputability of Y by reducing computing X to computing Y. All of these results except for the MRDP Theorem appear in either the text or exercises. The Halting Problem serves as our starting point for all these uncomputability results as well as many others.

The idea behind such uncomputability results is conceptually simple but can at first be quite confusing. If we know that is uncomputable, and we want to show that some other function is uncomputable, then we can do so via a *contrapositive* argument (i.e., proof by contradiction). That is, we show that **if** there exists a Turing machine that computes **then** there exists a Turing machine that computes . (Indeed, this is exactly how we showed that itself is uncomputable, by reducing this fact to the uncomputability of the function from uncomputable-func.)

For example, to prove that is uncomputable, we could show that there is a computable function such that for every pair and , . The existence of such a function implies that **if** was computable **then** would be computable as well, hence leading to a contradiction! The confusing part about reductions is that we are assuming something we *believe* is false (that has an algorithm) to derive something that we *know* is false (that has an algorithm). Michael Sipser describes such results as having the form *“If pigs could whistle then horses could fly”*.

A reduction-based proof has two components. For starters, since we need to be computable, we should describe the algorithm to compute it. The algorithm to compute is known as a *reduction* since the transformation modifies an input to to an input to , and hence *reduces* the task of computing to the task of computing . The second component of a reduction-based proof is the *analysis* of the algorithm : namely a proof that does indeed satisfy the desired properties.

Reduction-based proofs are just like other proofs by contradiction, but the fact that they involve hypothetical algorithms that don’t really exist tends to make reductions quite confusing. The one silver lining is that at the end of the day the notion of reductions is mathematically quite simple, and so it’s not that bad even if you have to go back to first principles every time you need to remember what is the direction that a reduction should go in.

A reduction is an *algorithm*, which means that, as discussed in implspecanarem, a reduction has three components:

* **Specification (what):** In the case of a reduction from to , the specification is that function should satisfy that for every Turing machine and input . In general, to reduce a function to , the reduction should satisfy for every input to .
* **Implementation (how):** The algorithm’s description: the precise instructions how to transform an input to the output .
* **Analysis (why):** A *proof* that the algorithm meets the specification. In particular, in a reduction from to this is a proof that for every input , the output of the algorithm satisfies that .

### Example: Halting on the zero problem

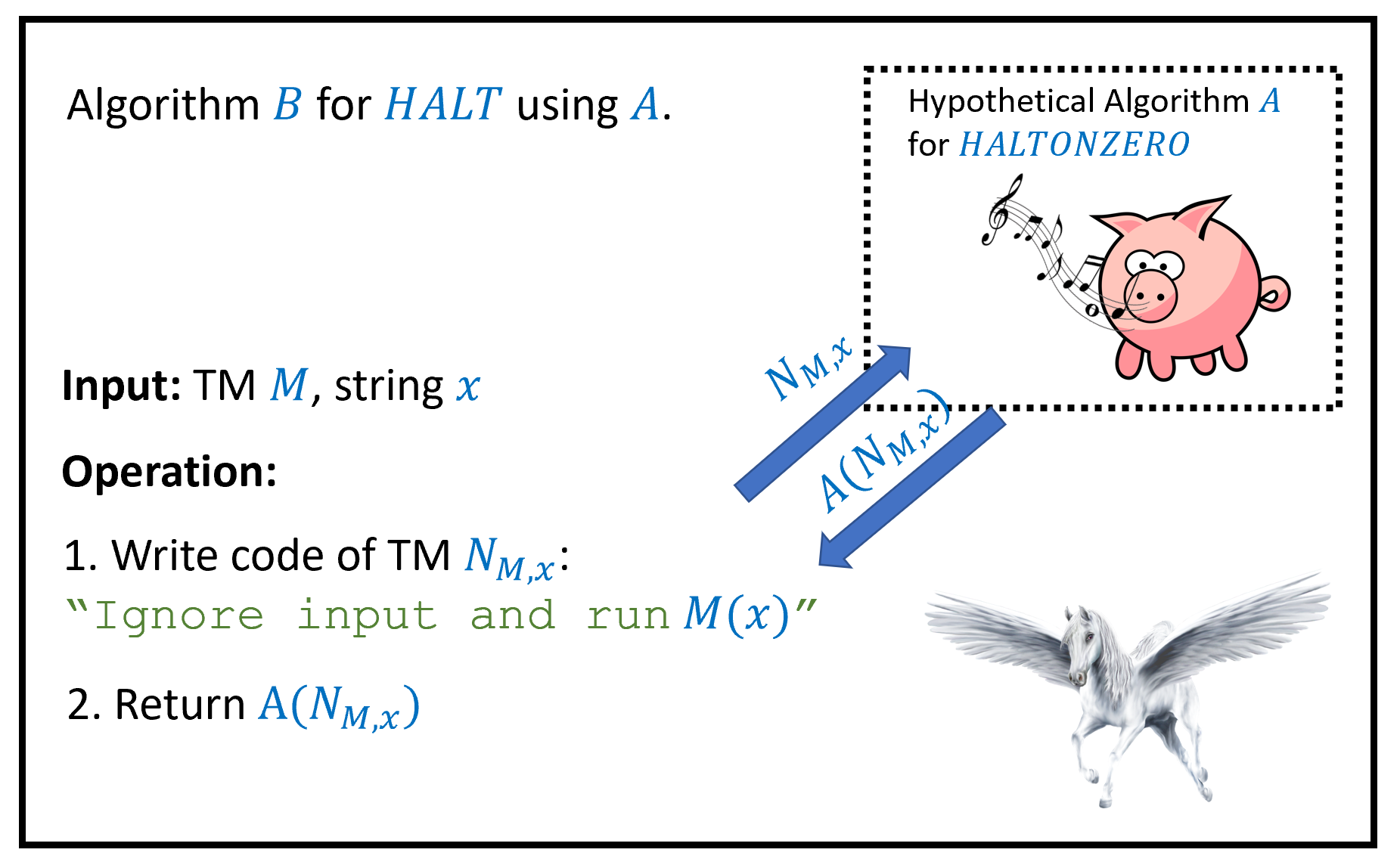
Here is a concrete example for a proof by reduction. We define the function as follows. Given any string , if and only if describes a Turing machine that halts when it is given the string as input. A priori seems like a potentially easier function to compute than the full-fledged function, and so we could perhaps hope that it is not uncomputable. Alas, the following theorem shows that this is not the case:

### 

is uncomputable.

### 

The proof of haltonzero-thm is below, but before reading it you might want to pause for a couple of minutes and think how you would prove it yourself. In particular, try to think of what a reduction from to would look like. Doing so is an excellent way to get some initial comfort with the notion of proofs by reduction, which a technique we will be using time and again in this book.



To prove haltonzero-thm, we show that is uncomputable by giving a *reduction* from the task of computing to the task of computing . This shows that if there was a hypothetical algorithm computing , then there would be an algorithm computing , contradicting halt-thm. Since neither nor actually exists, this is an example of an implication of the form “if pigs could whistle then horses could fly”.

The proof is by reduction from , see haltonzerofig. We will assume, towards the sake of contradiction, that is computable by some algorithm , and use this hypothetical algorithm to construct an algorithm to compute , hence obtaining a contradiction to halt-thm. (As discussed in defalgsec, following our “eat your cake and have it too” paradigm, we just use the generic name “algorithm” rather than worrying whether we model them as Turing machines, NAND-TM programs, NAND-RAM, etc.; this makes no difference since all these models are equivalent to one another.)

Since this is our first proof by reduction from the Halting problem, we will spell it out in more details than usual. Such a proof by reduction consists of two steps:

1. *Description of the reduction:* We will describe the operation of our algorithm , and how it makes “function calls” to the hypothetical algorithm .
2. *Analysis of the reduction:* We will then prove that under the hypothesis that Algorithm computes , Algorithm will compute .

INPUT: Turing machine $M$ and string $x$.  
OUTPUT: Turing machine $M'$ such that $M$ halts on $x$ iff $M'$ halts on zero  
  
Procedure{$N\_{M,x}$}{$w$} # Description of the T.M. $N\_{M,x}$  
 Return $EVAL(M,x)$ # Ignore the input $w$, evaluate $M$ on $x$.  
Endprocedure  
  
Return $N\_{M,x}$ # We do not execute $N\_{M,x}$: only return its description

Our Algorithm works as follows: on input , it runs halttohaltonzerored to obtain a Turing Machine , and then returns . The machine ignores its input and simply runs on .

In pseudocode, the program will look something like the following:

def N(z):  
 M = r'.......'  
 # a string constant containing desc. of M  
 x = r'.......'  
 # a string constant containing x  
 return eval(M,x)  
 # note that we ignore the input z

That is, if we think of as a program, then it is a program that contains and as “hardwired constants”, and given any input , it simply ignores the input and always returns the result of evaluating on . The algorithm does *not* actually execute the machine . merely writes down the description of as a string (just as we did above) and feeds this string as input to .

The above completes the *description* of the reduction. The *analysis* is obtained by proving the following claim:

**Claim:** For every strings , the machine constructed by Algorithm in Step 1 satisfies that halts on if and only if the program described by halts on the input .

**Proof of Claim:** Since ignores its input and evaluates on using the universal Turing machine, it will halt on if and only if halts on .

In particular if we instantiate this claim with the input to , we see that . Thus if the hypothetical algorithm satisfies for every then the algorithm we construct satisfies for every , contradicting the uncomputability of .

### 

In the proof of haltonzero-thm we used the technique of “hardwiring” an input to a program/machine . That is, modifying a program that it uses “hardwired constants” for some of all of its input. This technique is quite common in reductions and elsewhere, and we will often use it again in this course.

## Rice’s Theorem and the impossibility of general software verification

The uncomputability of the Halting problem turns out to be a special case of a much more general phenomenon. Namely, that *we cannot certify semantic properties of general purpose programs*. “Semantic properties” mean properties of the *function* that the program computes, as opposed to properties that depend on the particular syntax used by the program.

An example for a *semantic property* of a program is the property that whenever is given an input string with an even number of ’s, it outputs . Another example is the property that will always halt whenever the input ends with a . In contrast, the property that a C program contains a comment before every function declaration is not a semantic property, since it depends on the actual source code as opposed to the input/output relation.

Checking semantic properties of programs is of great interest, as it corresponds to checking whether a program conforms to a specification. Alas it turns out that such properties are in general *uncomputable*. We have already seen some examples of uncomputable semantic functions, namely and , but these are just the “tip of the iceberg”. We start by observing one more such example:

### 

Let be the function such that for every , if and only if represents a Turing machine such that outputs on every input . Then is uncomputable.

Despite the similarity in their names, and are two different functions. For example, if is a Turing machine that on input , halts and outputs the OR of all of ’s coordinates, then (since does halt on the input ) but (since does not compute the constant zero function).

The proof is by reduction to . Suppose, towards the sake of contradiction, that there was an algorithm such that for every . Then we will construct an algorithm that solves , contradicting haltonzero-thm.

Given a Turing machine (which is the input to ), our Algorithm does the following:

1. Construct a Turing Machine which on input , first runs and then outputs .
2. Return .

Now if halts on the input then the Turing machine computes the constant zero function, and hence under our assumption that computes , . If does not halt on the input , then the Turing machine will not halt on any input, and so in particular will *not* compute the constant zero function. Hence under our assumption that computes , . We see that in both cases, and hence the value that Algorithm returns in step 2 is equal to which is what we needed to prove.

Another result along similar lines is the following:

### 

The following function is uncomputable

We leave the proof of paritythm as an exercise (paritythmex). I strongly encourage you to stop here and try to solve this exercise.

### Rice’s Theorem

paritythm can be generalized far beyond the parity function. In fact, this generalization rules out verifying any type of semantic specification on programs. We define a *semantic specification* on programs to be some property that does not depend on the code of the program but just on the function that the program computes.

For example, consider the following two C programs

int First(int k) {  
 return 2\*k;  
}

int Second(int n) {  
 int i = 0;  
 int j = 0  
 while (j<n) {  
 i = i + 2;  
 j= j + 1;  
 }  
 return i;  
}

First and Second are two distinct C programs, but they compute the same function. A *semantic* property, would be either *true* for both programs or *false* for both programs, since it depends on the *function* the programs compute and not on their code. An example for a semantic property that both First and Second satisfy is the following: *“The program computes a function mapping integers to integers satisfying that for every input ”.*

A property is *not semantic* if it depends on the *source code* rather than the input/output behavior. For example, properties such as “the program contains the variable k” or “the program uses the while operation” are not semantic. Such properties can be true for one of the programs and false for others. Formally, we define semantic properties as follows:

A pair of Turing machines and are *functionally equivalent* if for every , . (In particular, iff for all .)

A function is *semantic* if for every pair of strings that represent functionally equivalent Turing machines, . (Recall that we assume that every string represents *some* Turing machine, see TMrepremark)

There are two trivial examples of semantic functions: the constant one function and the constant zero function. For example, if is the constant zero function (i.e., for every ) then clearly for every pair of Turing machines and that are functionally equivalent and . Here is a non-trivial example

Prove that the function is semantic.

Recall that if and only if for every . If and are functionally equivalent, then for every , . Hence if and only if .

Often the properties of programs that we are most interested in computing are the *semantic* ones, since we want to understand the programs’ functionality. Unfortunately, Rice’s Theorem tells us that these properties are all uncomputable:

Let . If is semantic and non-trivial then it is uncomputable.

The idea behind the proof is to show that every semantic non-trivial function is at least as hard to compute as . This will conclude the proof since by haltonzero-thm, is uncomputable. If a function is non trivial then there are two machines and such that and . So, the goal would be to take a machine and find a way to map it into a machine , such that **(i)** if halts on zero then is functionally equivalent to and **(ii)** if does *not* halt on zero then is functionally equivalent .

Because is semantic, if we achieved this, then we would be guaranteed that , and hence would show that if was computable, then would be computable as well, contradicting haltonzero-thm.

We will not give the proof in full formality, but rather illustrate the proof idea by restricting our attention to a particular semantic function . However, the same techniques generalize to all possible semantic functions. Define as follows: if there does not exist and two inputs such that for every but outputs and . That is, if it’s not possible to find an input such that flipping some bits of from to will change ’s output in the other direction from to . We will prove that is uncomputable, but the proof will easily generalize to any semantic function.

We start by noting that is neither the constant zero nor the constant one function:

* The machine that simply goes into an infinite loop on every input satisfies , since is not defined *anywhere* and so in particular there are no two inputs where for every but and .
* The machine that computes the XOR or parity of its input, is not monotone (e.g., but ) and hence .

(Note that and are *machines* and not *functions*.)

We will now give a reduction from to . That is, we assume towards a contradiction that there exists an algorithm that computes and we will build an algorithm that computes . Our algorithm will work as follows:

**Algorithm :**

**Input:** String describing a Turing machine. (*Goal:* Compute )

**Assumption:** Access to Algorithm to compute .

**Operation:**

1. Construct the following machine : “On input do: **(a)** Run , **(b)** Return ”.
2. Return .

To complete the proof we need to show that outputs the correct answer, under our assumption that computes . In other words, we need to show that . Suppose that does *not* halt on zero. In this case the program constructed by Algorithm enters into an infinite loop in step **(a)** and will never reach step **(b)**. Hence in this case is functionally equivalent to . (The machine is not the same machine as : its description or *code* is different. But it does have the same input/output behavior (in this case) of never halting on any input. Also, while the program will go into an infinite loop on every input, Algorithm never actually runs : it only produces its code and feeds it to . Hence Algorithm will *not* enter into an infinite loop even in this case.) Thus in this case, .

If *does* halt on zero, then step **(a)** in will eventually conclude and ’s output will be determined by step **(b)**, where it simply outputs the parity of its input. Hence in this case, computes the non-monotone parity function (i.e., is functionally equivalent to ), and so we get that . In both cases, , which is what we wanted to prove.

An examination of this proof shows that we did not use anything about beyond the fact that it is semantic and non-trivial. For every semantic non-trivial , we can use the same proof, replacing and with two machines and such that and . Such machines must exist if is non trivial.

Rice’s Theorem is so powerful and such a popular way of proving uncomputability that people sometimes get confused and think that it is the *only* way to prove uncomputability. In particular, a common misconception is that if a function is *not* semantic then it is *computable*. This is not at all the case.

For example, consider the following function . This is a function that on input a string that represents a NAND-TM program , outputs if and only if both **(i)** halts on the input , and **(ii)** the program does not contain a variable with the identifier Yale. The function is clearly not semantic, as it will output two different values when given as input one of the following two functionally equivalent programs:

Yale[0] = NAND(X[0],X[0])  
Y[0] = NAND(X[0],Yale[0])

and

Harvard[0] = NAND(X[0],X[0])  
Y[0] = NAND(X[0],Harvard[0])

However, is uncomputable since every program can be transformed into an equivalent (and in fact improved :)) program that does not contain the variable Yale. Hence if we could compute then determine halting on zero for NAND-TM programs (and hence for Turing machines as well).

Moreover, as we will see in godelchap, there are uncomputable functions whose inputs are not programs, and hence for which the adjective “semantic” is not applicable.

Properties such as “the program contains the variable Yale” are sometimes known as *syntactic* properties. The terms “semantic” and “syntactic” are used beyond the realm of programming languages: a famous example of a syntactically correct but semantically meaningless sentence in English is Chomsky’s [“Colorless green ideas sleep furiously.”](https://goo.gl/4gXoiV) However, formally defining “syntactic properties” is rather subtle and we will not use this terminology in this book, sticking to the terms “semantic” and “non semantic” only.

### Halting and Rice’s Theorem for other Turing-complete models

As we saw before, many natural computational models turn out to be *equivalent* to one another, in the sense that we can transform a “program” of one model (such as a expression, or a game-of-life configurations) into another model (such as a NAND-TM program). This equivalence implies that we can translate the uncomputability of the Halting problem for NAND-TM programs into uncomputability for Halting in other models. For example:

### 

Let be the function that on input strings and outputs if the NAND-TM program described by halts on the input and outputs otherwise. Then is uncomputable.

### 

Once again, this is a good point for you to stop and try to prove the result yourself before reading the proof below.

We have seen in TM-equiv-thm that for every Turing machine , there is an equivalent NAND-TM program such that for every , . In particular this means that .

The transformation that is obtained from the proof of TM-equiv-thm is *constructive*. That is, the proof yields a way to *compute* the map . This means that this proof yields a *reduction* from task of computing to the task of computing , which means that since is uncomputable, neither is .

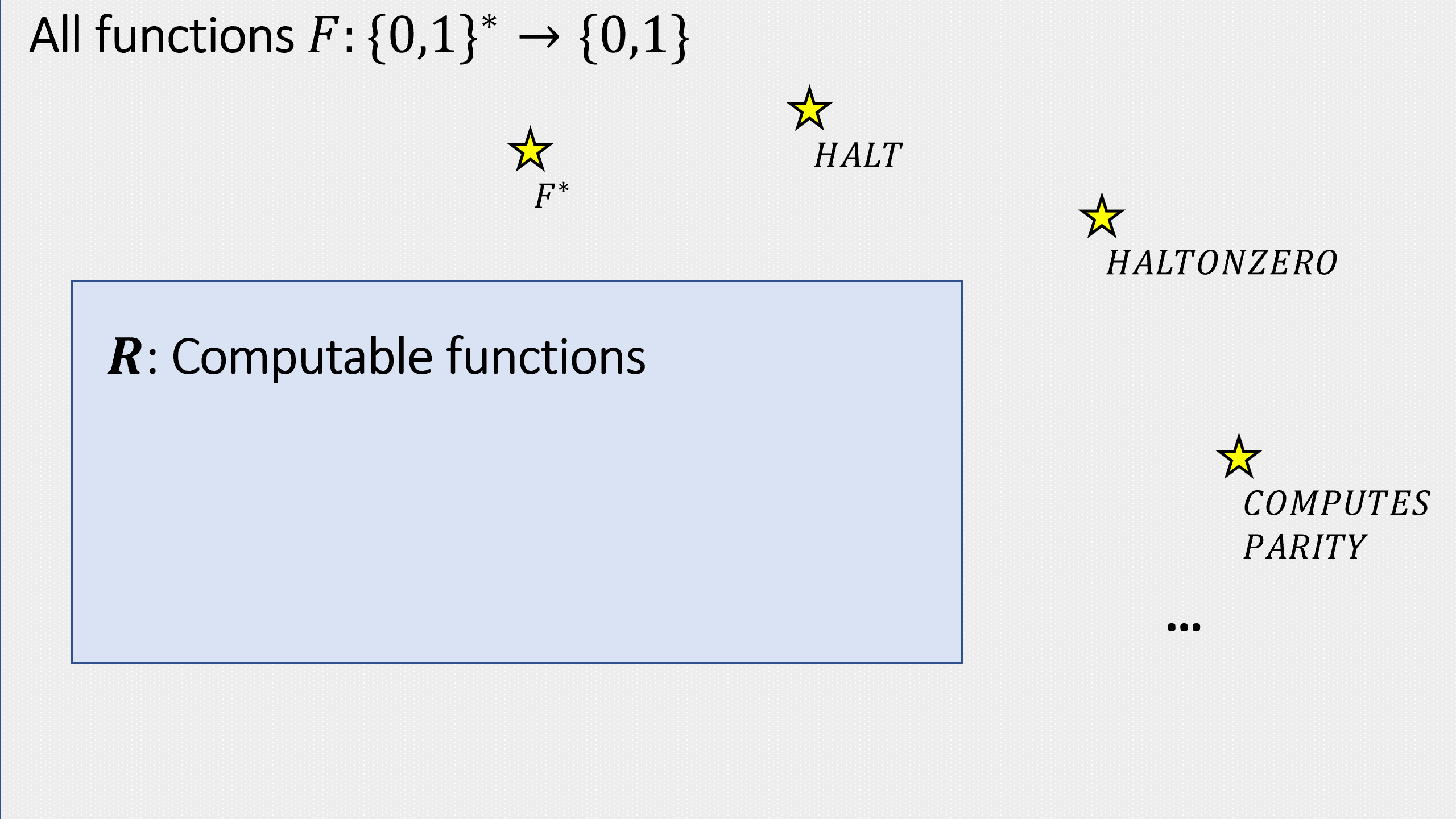
The same proof carries over to other computational models such as the  *calculus*, *two dimensional* (or even one-dimensional) *automata* etc. Hence for example, there is no algorithm to decide if a expression evaluates the identity function, and no algorithm to decide whether an initial configuration of the game of life will result in eventually coloring the cell black or not.

Indeed, we can generalize Rice’s Theorem to all these models. For example, if is a non-trivial function such that for every functionally equivalent NAND-TM programs then is uncomputable, and the same holds for NAND-RAM programs, -expressions, and all other Turing complete models (as defined in turingcompletedef), see also ricegeneralex.

### Is software verification doomed? (discussion)

Programs are increasingly being used for mission critical purposes, whether it’s running our banking system, flying planes, or monitoring nuclear reactors. If we can’t even give a certification algorithm that a program correctly computes the parity function, how can we ever be assured that a program does what it is supposed to do? The key insight is that while it is impossible to certify that a *general* program conforms with a specification, it is possible to write a program in the first place in a way that will make it easier to certify. As a trivial example, if you write a program without loops, then you can certify that it halts. Also, while it might not be possible to certify that an *arbitrary* program computes the parity function, it is quite possible to write a particular program for which we can mathematically *prove* that computes the parity. In fact, writing programs or algorithms and providing proofs for their correctness is what we do all the time in algorithms research.

The field of *software verification* is concerned with verifying that given programs satisfy certain conditions. These conditions can be that the program computes a certain function, that it never writes into a dangerous memory location, that is respects certain invariants, and others. While the general tasks of verifying this may be uncomputable, researchers have managed to do so for many interesting cases, especially if the program is written in the first place in a formalism or programming language that makes verification easier. That said, verification, especially of large and complex programs, remains a highly challenging task in practice as well, and the number of programs that have been formally proven correct is still quite small. Moreover, even phrasing the right theorem to prove (i.e., the specification) if often a highly non-trivial endeavor.



The set of computable Boolean functions (classRdef) is a proper subset of the set of all functions mapping to . In this chapter we saw a few examples of elements in the latter set that are not in the former.

* There is a *universal* Turing machine (or NAND-TM program) such that on input a description of a Turing machine and some input , halts and outputs if (and only if) halts on input . Unlike in the case of finite computation (i.e., NAND-CIRC programs / circuits), the input to the program can be a machine that has more states than itself.
* Unlike the finite case, there are actually functions that are *inherently uncomputable* in the sense that they cannot be computed by *any* Turing machine.
* These include not only some “degenerate” or “esoteric” functions but also functions that people have deeply care about and conjectured that could be computed.
* If the Church-Turing thesis holds then a function that is uncomputable according to our definition cannot be computed by any means in our physical world.

## Exercises

Let be the function such that on input where represents a NAND-RAM program, iff halts on the input . Prove that is uncomputable.

Let be the function that on input (a string representing) a triple , iff the Turing machine , on input , halts within at most steps (where a *step* is defined as one sequence of reading a symbol from the tape, updating the state, writing a new symbol and (potentially) moving the head).

Prove that is *computable*.

Let be the function that on input (a string representing) a triple , iff the Turing machine , on input , halts before its head reached the -th location of its tape. (We don’t care how many steps makes, as long as the head stays inside locations .)

Prove that is *computable*. See footnote for hint[[2]](#footnote-88)

Suppose that and are computable functions. For each one of the following functions , either prove that is *necessarily computable* or give an example of a pair and of computable functions such that will not be computable. Prove your assertions.

1. iff OR .
2. iff there exist two nonempty strings such that (i.e., is the concatenation of and ), and .
3. iff there exist a list of non empty strings such that strings for every and .
4. iff is a valid string representation of a NAND++ program such that for every , on input the program outputs .
5. iff is a valid string representation of a NAND++ program such that on input the program outputs .
6. iff is a valid string representation of a NAND++ program such that on input , outputs after executing at most lines.

Prove that the following function is uncomputable. On input , we define if and only if is a string that represents a NAND++ program such that there only a finite number of inputs s.t. .[[3]](#footnote-91)

Prove paritythm without using Rice’s Theorem.

Let be the function defined as follows: given a string representing a pair of Turing machines, iff and are functionally equivalent as per semanticpropdef. Prove that is uncomputable.

Note that you *cannot* use Rice’s Theorem directly, as this theorem only deals with functions that take a single Turing machine as input, and takes two machines.

For each of the following two functions, say whether it is computable or not:

1. Given a NAND-TM program , an input , and a number , when we run on , does the index variable i ever reach ?
2. Given a NAND-TM program , an input , and a number , when we run on , does ever write to an array at index ?

Let be the function that is defined as follows. On input a string that represents a NAND-RAM program and a String that represents a Turing machine, if and only if there exists some input such halts on but does not halt on . Prove that is uncomputable. See footnote for hint.[[4]](#footnote-96)

Define a function to be *recursively enumerable* if there exists a Turing machine such that such that for every , if then , and if then . (i.e., if then does not halt on .)

1. Prove that every computable is also recursively enumerable.
2. Prove that there exists that is not computable but is recursively enumerable. See footnote for hint.[[5]](#footnote-98)
3. Prove that there exists a function such that is not recursively enumerable. See footnote for hint.[[6]](#footnote-99)
4. Prove that there exists a function such that is recursively enumerable but the function defined as is *not* recursively enumerable. See footnote for hint.[[7]](#footnote-100)

In this exercise we will prove Rice’s Theorem in the form that it is typically stated in the literature.

For a Turing machine , define to be the set of all such that halts on the input and outputs . (The set is known in the literature as the *language recognized by* . Note that might either output a value other than or not halt at all on inputs . )

1. Prove that for every Turing Machine , if we define to be the function such that iff then is *recursively enumerable* as defined in recursiveenumerableex.
2. Use rice-thm to prove that for every , if **(a)** is neither the constant zero nor the constant one function, and **(b)** for every such that , , then is uncomputable. See footnote for hint.[[8]](#footnote-102)

Let be the set of all partial functions from to and be a Turing-equivalent model as defined in turingcompletedef. We define a function to be *-semantic* if there exists some such that for every .

Prove that for every -semantic that is neither the constant one nor the constant zero function, is uncomputable.

## Bibliographical notes

The cartoon of the Halting problem in universalchapoverviewfig and taken from [Charles Cooper’s website](https://www.coopertoons.com/education/haltingproblem/haltingproblem.html/).

Section 7.2 in [@MooreMertens11] gives a highly recommended overview of uncomputability. Gödel, Escher, Bach [@hofstadter1999] is a classic popular science book that touches on uncomputability, and unprovability, and specifically Gödel’s Theorem that we will see in godelchap. See also the recent book by Holt [@Holt2018].

The history of the definition of a function is intertwined with the development of mathematics as a field. For many years, a function was identified (as per Euler’s quote above) with the means to calculate the output from the input. In the 1800’s, with the invention of the Fourier series and with the systematic study of continuity and differentiability, people have started looking at more general kinds of functions, but the modern definition of a function as an arbitrary mapping was not yet universally accepted. For example, in 1899 Poincare wrote *“we have seen a mass of bizarre functions which appear to be forced to resemble as little as possible honest functions which serve some purpose. … they are invented on purpose to show that our ancestor’s reasoning was at fault, and we shall never get anything more than that out of them”.* Some of this fascinating history is discussed in [@grabiner1983gave, @Kleiner91, @Lutzen2002, @grabiner2005the ].

The existence of a universal Turing machine, and the uncomputability of was first shown by Turing in his seminal paper [@Turing37], though closely related results were shown by Church a year before. These works built on Gödel’s 1931 *incompleteness theorem* that we will discuss in godelchap.

Some universal Turing Machines with a small alphabet and number of states are given in [@rogozhin1996small], including a single-tape universal Turing machine with the binary alphabet and with less than states; see also the survey [@woods2009complexity]. Adam Yedidia has written [software](https://github.com/adamyedidia/parsimony) to help in producing Turing machines with a small number of states. This is related to the recreational pastime of [“Code Golfing”](https://codegolf.stackexchange.com/) which is about solving a certain computational task using the as short as possible program.

The diagonalization argument used to prove uncomputability of is derived from Cantor’s argument for the uncountability of the reals discussed in chaprepres.

Christopher Strachey was an English computer scientist and the inventor of the CPL programming language. He was also an early artificial intelligence visionary, programming a computer to play Checkers and even write love letters in the early 1950’s, see [this New Yorker article](https://www.newyorker.com/tech/elements/christopher-stracheys-nineteen-fifties-love-machine) and [this website](http://www.alpha60.de/art/love_letters/).

Rice’s Theorem was proven in [@rice1953classes]. It is typically stated in a form somewhat different than what we used, see ricesstandardex.

We do not discuss in the chapter the concept of *recursively enumerable* languages, but it is covered briefly in recursiveenumerableex. As usual, we use function, as opposto language, notation.

The cartoon of the Halting problem in universalchapoverviewfig is copyright 2019 Charles F. Cooper.

1. This argument has also been connected to the issues of consciousness and free will. I am personally skeptical of its relevance to these issues. Perhaps the reasoning is that humans have the ability to solve the halting problem but they exercise their free will and consciousness by choosing not to do so. [↑](#footnote-ref-53)
2. A machine with alphabet can have at most choices for the contents of the first locations of its tape. What happens if the machine repeats a previously seen configuration, in the sense that the tape contents, the head location, and the current state, are all identical to what they were in some previous state of the execution? [↑](#footnote-ref-88)
3. Hint: You can use Rice’s Theorem. [↑](#footnote-ref-91)
4. *Hint:* While it cannot be applied directly, with a little “massaging” you can prove this using Rice’s Theorem. [↑](#footnote-ref-96)
5. has this property. [↑](#footnote-ref-98)
6. You can either use the diagonalization method to prove this directly or show that the set of all recursively enumerable functions is *countable*. [↑](#footnote-ref-99)
7. has this property: show that if both and were recursively enumerable then would be in fact computable. [↑](#footnote-ref-100)
8. Show that any satisfying **(b)** must be semantic. [↑](#footnote-ref-102)