Formal methods: a plea for knowing what computers do

Gilles Dowek

The Nasa formal method team is 30 years old

This talk: formal methods in the next 30 years

We shall need more formal methods because we shall need more and more to understand what computers do

I. A paradoxical wish

Computers have been invented to surprise us

Why do compute surprise us?

Would you bet 8 to 1 that the thousandth digit of π is a 9?

Why do compute surprise us?

Would you bet 8 to 1 that the thousandth digit of π is a 9?

It is

Why do compute surprise us?

If we knew what computers do, we would not use them

and we would not have built any

But a safety concern

Can we trust a black box system if we do not know what is does?

And also an ethical concern

- A piece of software rejects your loan application, gives you a bad grade, imposes you a fine..., you have the right to know why
- Not only the law must be public, but you are not supposed to ignore it (and you have the right to "manipulate" the law : example drive at 54.9 when the speed limit is 55 m/h)
- Requiring an explanation is a defence against arbitrary decisions (e.g. loan application)
- Requiring an explanation is a defence against ethnic biases (an explanation: "the loan has been refused because you have a blue (as opposed to green) skin")
- An explanation is a way to progress (e.g. grading an exam)

A paradox

Explanation (safety, fairness...): a higher concern concern when a computer makes the decision

- The Maître D' paradox

- "I'd rather fly in an aircraft with a pilot and a probability of collision of 10^{-6} than in an aircraft with no pilot and a probability of collision of 10^{-7}"
- An autonomous car runs over a pedestrian

- Amazon does not deliver in poor areas, but gourmet restaurants are not in poor areas either
- Judges before and after lunch

II. Specifying and verifying

Knowing what computers do

is not being able to predict the result
is not being able to know how the result has been computed

It is being able to predict a property of the result

- it is the thousandth digits of π
- Air traffic control: no conflicts, no collisions will happen
- Loan application: the result does not depend on skin color
- ...

Writing the specification

is already difficult, but useful

Ask yourself what you expect of your program

- No conflicts? But also the aircraft to reach destination (eventually, in less that 24 hours...?)
- No discrimination: but what is discrimination?: your loan application has been rejected because you are a woman / because you are five year old

Avoiding errors and avoiding malevolence

Safety and ethics:
two reasons to wish to know what computer do

From certified to certifying

Certified and certifying compilers

Certified: what the software will do on any input
Certifying: what the software has done on a specific input

Certificates: a path explaining why to points are connected in a graph, a divisor explaining why a number is composite

If a system is certified then it is certifying
If it is certifying, it is almost certified (what is a proof of $\forall x\ A$?)

From certified to certifying

People may be certifying, never certified

Automatic bankers, judges, teachers... should be certifying
Algorithms for autonomous vehicules should be certified

III. Two cases where explaining is challenging

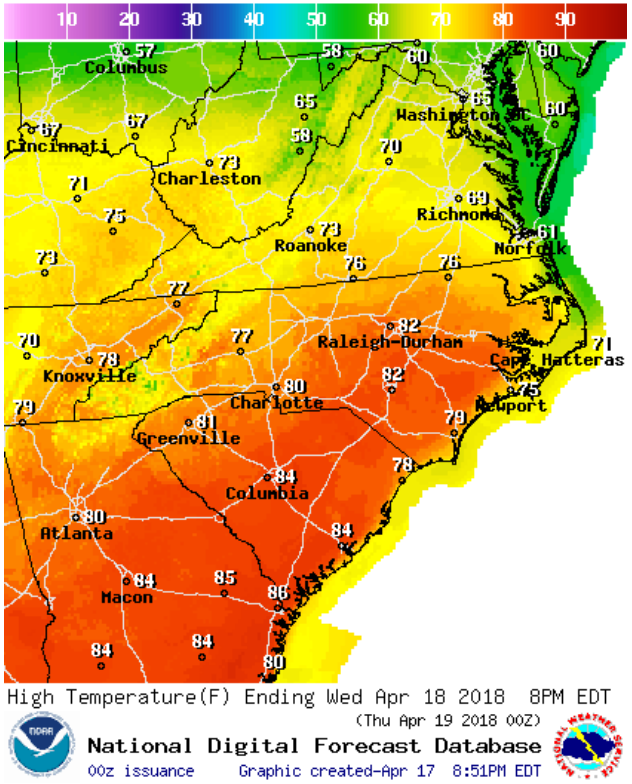
A computation that takes several hours

Billions of operations

Each of them easy to explain

But putting together one billion of explanations does not yield an explanation

Weather forecast



Tomorrow temperature in Newport News: 61°F

Why 61 and not 60 or 62?

What does it mean to explain 61?

Not predicting the result, not explaining each line of code...

El niño moved in this or that direction?

From weather to climate

Weather programs can be black boxes
But models of climate should not

Learning algorithms

Edward Hopper

Keith Haring

Keith Haring

Edward Hopper

Edward Hopper or Keith Haring?

Can you explain why you said that?

Image #5 is closer to image #2 and #3 than to image #1 and #4

A distance between images
A base of labeled images
Compute average distance to each cluster, pick the smallest

What is there to explain?

The colors are “brighter” in Haring’s painting than in Hopper’s

May be brightness is part of the definition of the distance:
explaining the algorithm is explaining the metric

Can you define “brightness”?

Moreover...

A multilayered algorithm defines its predicates
May be not "brightness" but "jadjegness"

Not always an explanation (that you can understand)

But in some cases...

Train an algorithm with judiciary cases where all the people with a blue skin are imposed a fine, and those who have a green skin not

This algorithm decides to impose a fine to a blue skin person

There is an explanation: she has been imposed a fine, because she has a blue skin

But...

No way to prove that the algorithm is not color-biased

Because the bias does does not come from the algorithm alone, but also on the data used to train it

The four concepts of Computer Science

Algorithm

Machine

Language

Data

The four concepts of Computer Science

Algorithm: we have specified and proved some correct

Machine

Language

Data

The four concepts of Computer Science

Algorithm: we have specified and proved some correct

Machine: we have specified and proved some correct

Language

Data

The four concepts of Computer Science

Algorithm: we have specified and proved some correct

Machine: we have specified and proved some correct

Language: we have specified and proved some correct

Data

The four concepts of Computer Science

Algorithm: we have specified and proved some correct

Machine: we have specified and proved some correct

Language: we have specified and proved some correct

Data:?

The future of formal methods

More formal methods because we shall need more explanations

Not only safety properties and security properties, but also ethical properties

Not only algorithms, machines, and languages proved correct but also data

A lot of work for the coming 30 years