Explanation: from ethics to logic

Gilles Dowek

My bank loan application has been rejected (by a clerk / a piece of software)

I want an explanation

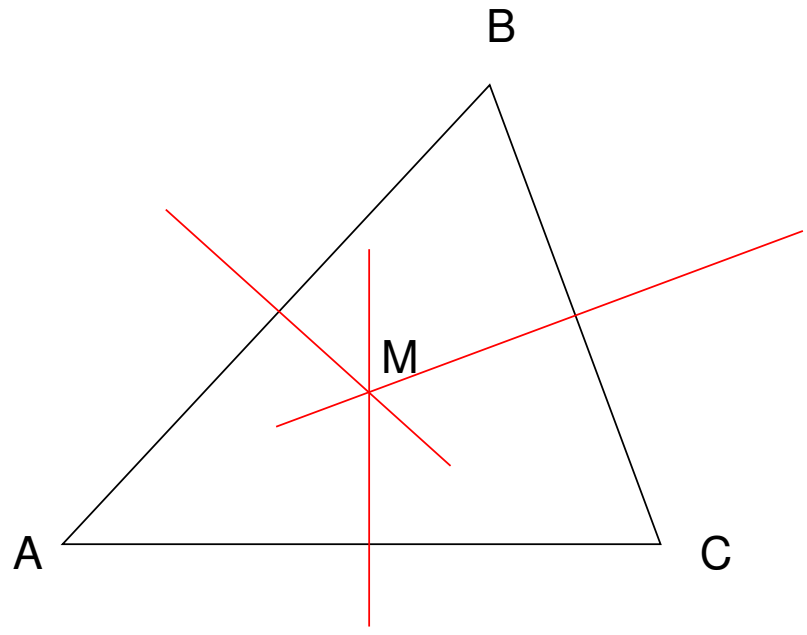
Transparency: I want to know the rules
Equality: I want to be sure the same rules apply to everyone
Agency: I want to be able to "improve" and apply again
Dignity: I want to be considered as a rational being

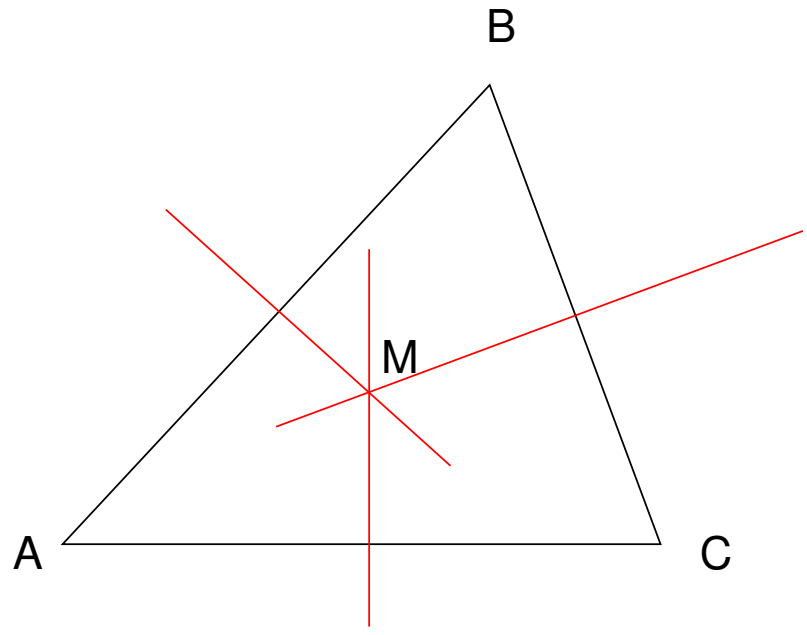
Explanation itself as a value?

I. Proof as an explanation

What is explained? A statement

What explains it? A proof of this statement


B
M
A
C

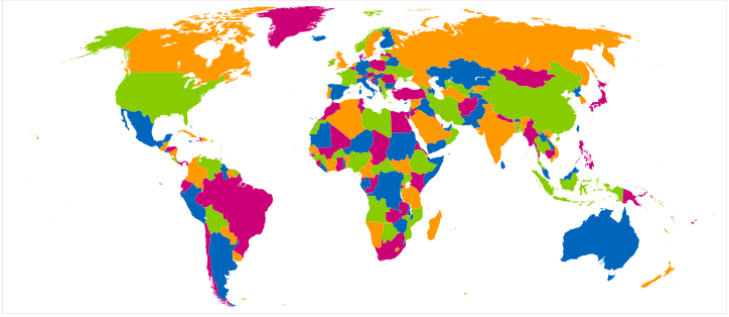


$d(M, A) = d(M, B)$ and $d(M, B) = d(M, C)$, hence
$d(M, A) = d(M, C)$

At the same time: that the bisectors are concurrent and why the bisectors are concurrent

Proof without explanation

1976: The four color theorem

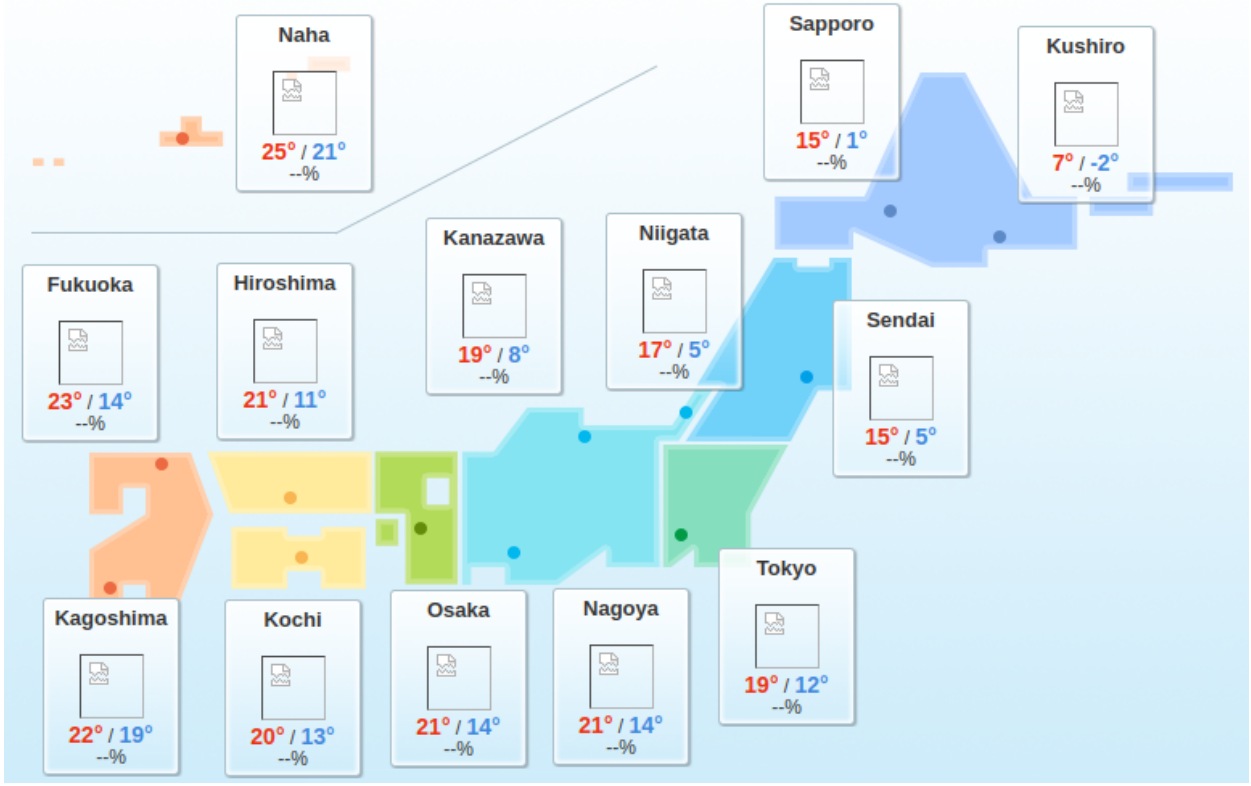


Every map is four colorable
Reduce the infinitude of possible maps to 1482
Check that these 1482 maps are "reducible": case study
"First" proof by computer

Rather calls for an explanation than provides one
Drawing 1482 triangles and checking that the bisectors are concurrent is puzzling rather than explanatory

Proof without explanation

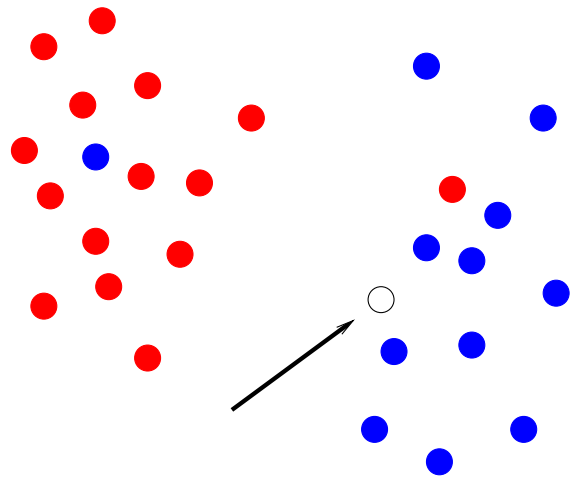
Weather forecast: a very long numerical analysis computation leading to the map



Why 19°C in Tokyo and not 18°C?

Proof without explanation

Data-centric algorithms (for instance, machine learning)



Cats and dogs

The new point is most likely a dog picture (average distance to dogs $<$ average distance to cats)

Why is it a dog? (Is it the size of her ears or the shape of her muzzle?)

Proof without explanation

Multiplication of arbitrary numbers

$$
\begin{array}{r}
7678 \\
3706 \\
\hline
46068 \\
00000 \\
53746 \\
23034 \\
\hline
28454668
\end{array}
$$

Why is the result of the multiplication 28454668?

Proof without explanation

These proofs (four color theorem, weather forecast, cats and dogs, multiplication) are not explanations

Explaining the results of machine learning algorithms is an issue, but the problem is not specific to machine learning (numerical analysis raises the same concern)

And sometimes difficult to tell what an explanation would look like

II. A first definition

$$\begin{array}{r} 1\,2\,3\,4\,5\,6\,7\,9 \\ 3\,6 \\ \hline \end{array}$$

$$\begin{array}{r} 1\,2\,3\,4\,5\,6\,7\,9 \\ 3\,6 \\ \hline 4 \end{array}$$

$$\begin{array}{r} 1\,2\,3\,4\,5\,6\,7\,9 \\ 3\,6 \\ \hline 7\,4 \end{array}$$

$$\begin{array}{r} 12345679 \\ 36 \\ \hline 074 \end{array}$$

$$\begin{array}{r} 12345679 \\ 36 \\ \hline 4074 \end{array}$$

$$\begin{array}{r} 12345679 \\ 36 \\ \hline 74074 \end{array}$$

$$\begin{array}{r} 12345679 \\ 36 \\ \hline 074074 \end{array}$$

$$\begin{array}{r} 12345679 \\ \times\quad 36 \\ \hline 4074074 \end{array}$$

$$\begin{array}{r} 12345679 \\ \times \quad 36 \\ \hline 74074074 \end{array}$$

$$\begin{array}{r} 12345679 \\ \times\quad 36 \\ \hline 740740747 \end{array}$$

$$
\begin{array}{r}
12345679 \\
\qquad\quad 36 \\
\hline
740740740 \\
\qquad\quad 37
\end{array}
$$

$$\begin{array}{r} 12345679 \\ \times\quad 36 \\ \hline 74074074 \\ 037 \end{array}$$

$$
\begin{array}{r}
12345679 \\
\qquad\quad 36 \\
\hline
740740740 \\
\quad 7037
\end{array}
$$

$$
\begin{array}{r}
12345679 \\
\qquad\quad 36 \\
\hline
74074074 \\
\quad 37037
\end{array}
$$

$$
\begin{array}{r}
12345679 \\
\times \quad 36 \\
\hline
74074074 \\
037037
\end{array}
$$

$$
\begin{array}{r}
12345679 \\
\qquad\quad 36 \\
\hline
74074074 \\
70370370
\end{array}
$$

$$
\begin{array}{r}
12345679 \\
\times \quad 36 \\
\hline
74074074 \\
370370370
\end{array}
$$

$$
\begin{array}{r}
12345679 \\
\qquad\quad 36 \\
\hline
74074074 \\
37037037 \\
\hline
\end{array}
$$

$$
\begin{array}{r}
12345679 \\
\qquad\quad 36 \\
\hline
74074074 \\
37037037 \\
\hline
\qquad\quad 4
\end{array}
$$

$$
\begin{array}{r}
12345679 \\
\qquad\quad 36 \\
\hline
74074074 \\
37037037 \\
\hline
\qquad\quad 44
\end{array}
$$

$$
\begin{array}{r}
12345679 \\
\qquad\qquad 36 \\
\hline
74074074 \\
37037037 \\
\hline
\qquad\qquad 444
\end{array}
$$

$$
\begin{array}{r}
12345679 \\
\qquad\qquad 36 \\
\hline
74074074 \\
37037037 \\
\hline
\qquad\qquad 4444
\end{array}
$$

$$
\begin{array}{r}
12345679 \\
\times \quad 36 \\
\hline
74074074 \\
370370370 \\
\hline
444444
\end{array}
$$

$$
\begin{array}{r}
12345679 \\
\times \quad 36 \\
\hline
74074074 \\
37037037 \\
\hline
444444
\end{array}
$$

$$
\begin{array}{r}
12345679 \\
\times \quad 36 \\
\hline
74074074 \\
370370370 \\
\hline
444444444
\end{array}
$$

$$
\begin{array}{r}
12345679 \\
\qquad\qquad 36 \\
\hline
74074074 \\
37037037 \\
\hline
444444444
\end{array}
$$

$$
\begin{array}{r}
12345679 \\
\qquad\qquad 36 \\
\hline
74074074 \\
37037037 \\
\hline
4444444444
\end{array}
$$

$$12345679 \times 36 = 444444444$$

Why is the result made of 4's only?

A first bit of explanation

36 has something to do with 4: the theorem rephrases

$$12345679 \times 9 \times 4 = 111111111 \times 4$$

and it is a consequence of

$$12345679 \times 9 = 111111111$$

This explanation yields a generalization (Lewis Carroll's magic trick)

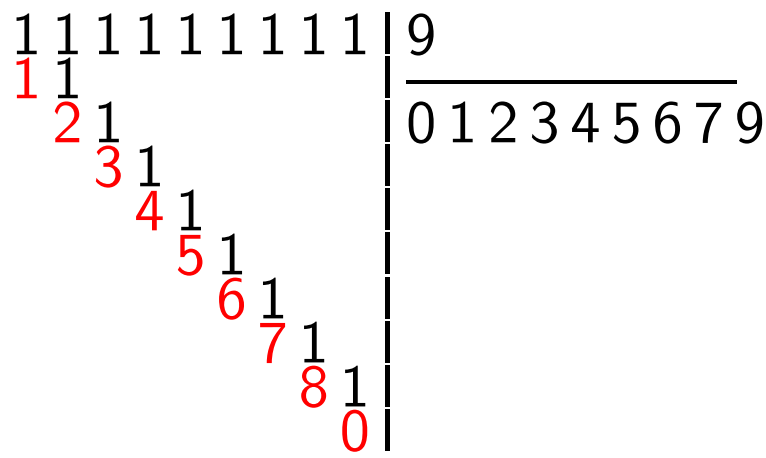
$$\forall n \in [1, 9]\ 12345679 \times 9 \times n = 111111111 \times n$$

For $n = 4$: $12345679 \times 36 = 444444444$
For $n = 7$: $12345679 \times 63 = 777777777$

But... that $12345679 \times 9 = 111111111$ remains to be explained

A deeper explanation

Rather explain why $111111111/9 = 12345679$



n-th digit of the result: $n-1$, also n-th partial remainder: n

By induction. $(n+1)$-th partial dividend: $10n+1 = 9n+n+1$

$(n+1)$-th digit of the result: n, $(n+1)$-th partial remainder: $n+1$

Works only when $n+1 < 9$, ~~the ninth digit is 8 and the ninth partial remainder is 9~~ the ninth digit is 9 and the ninth partial remainder is 0

A deeper explanation yields a wider generalization

Add more 1's until the next 0 partial remainder
$1111111111111111/9 = 12345679012345679$

$$12345679012345679 \times 36 = 4444444444444444$$

Works in any base, for instance base 20 with digits $(0, ..., 9, a, ..., j)$
$1111111111111111/j = 123456789abcdefghj$

$$123456789abcdefghj \times 3g = 4444444444444444$$

Explanation: a first definition

Instead of proving $12345679 \times 36 = 444444444$ with the multiplication algorithm, proved the generalization

$$\forall n \in [1, 9]\ 12345679 \times 9 \times n = 111111111 \times n$$

(in a generic way, that is without enumerating the nine cases) and deduced the truth of $12345679 \times 36 = 444444444$

Explanation: judgement of the truth of a particular statement proceeding by a judgement of the truth of a more general statement, and then a specialization to the particular case

The more general the intermediate statement, the deeper the explanation

Cut

A notion that already exists in proof theory: the notion of cut

$$\dfrac{\dfrac{\dfrac{\dots}{A[x]}}{\forall x\ A[x]}\ \forall\text{-introduction}}{A[t]}\ \forall\text{-elimination}$$

An explanation is a truth judgement through a proof with a cut

Revisiting the examples

- Bisectors of a triangle: works for every triangle, therefore for this one
- Bank loan: general statement "all applications without a guarantor are rejected", therefore...

- Arbitrary multiplication: difficult to find a general statement, hence difficult to find an explanation
- Weather forecast: difficult to find a general statement, hence difficult to find an explanation
- Data-centric algorithms: difficult to find a general statement, hence difficult to find an explanation, but dalmatians vs. dachshunds (explanatory AI)

Revisiting the examples

- Four color theorem: no generic proof: a different proof for each of the 1482 maps

Compare with the proof of

$$\forall n \in [1, 9]\ 12345679 \times 9 \times n = 111111111 \times n$$

that proves $12345679 \times 9 = 111111111$ and generically multiplies the two sides by n

Quantifier elimination, enumeration, and non-explanation

$$\exists x \ (x^2 = 1800)$$

Quantifier elimination, enumeration, and non-explanation

$$\exists x\ (x^2 = 1800)$$

$42^2 = 1764$
$43^2 = 1849$

Quantifier elimination, enumeration, and non-explanation

$$\exists x\ (x^2 = 1800)$$

$42^2 = 1764$
$43^2 = 1849$

$$\forall x\ (x \leq 42 \Rightarrow x^2 < 1800)$$

$$\forall x\ (x > 42 \Rightarrow x^2 > 1800)$$

Finite domain: generic proof or proof by enumeration
Infinite domain: generic proof
If both proofs generic: an explanation

Quantifier elimination, enumeration, and non-explanation

Generalizes to any univariate Diophantine equation

$$\exists x \ (a_n x^n + a_{n-1} x^{n-1} + ... + a_1 x + a_0 = 0)$$

$b = n \max(|a_0|, ..., |a_{n-1}|)$

$$\forall x \ (x > b \Rightarrow |a_n x^n| > |a_{n-1} x^{n-1} + ... + a_1 x + a_0|)$$

generic proof
Thus equivalent to

$$\exists x \ (x \leq b \wedge a_n x^n + a_{n-1} x^{n-1} + ... + a_1 x + a_0 = 0)$$

In general no generic refutation, but enumeration (finite domain)
Infinitude of natural numbers reduced to $b + 1$, checked one by one
Quantifier elimination (for example, the four color theorem)

III. From proof to algorithm

What is a proof of a general statement?

Brouwer-Heyting-Kolmogorov: a proof of $\forall x\ A[x]$ is an algorithm that maps each term t to a proof of $A[t]$

For example, a proof of

$$\forall n \in [1,9]\ 12345679 \times 9 \times n = 111111111 \times n$$

is an algorithm mapping every number p in $[1,9]$ to a proof of

$$12345679 \times 9 \times p = 111111111 \times p$$

A more general definition of the notion of explanation

If f is an algorithm and $\forall x\ (P(x) \Rightarrow Q(f(x)))$ a specification of this algorithm, then the pair formed with the algorithm f and the input value a is an explanation of $Q(b)$ where b is the output value $f(a)$

The postman delivered (Q) a book (b) this morning. What is the explanation of $Q(b)$? It is that I have placed (P) an order (a) to an on-line book shop (f) and $b = f(a)$

f: the book shop (algorithm mapping orders to books)
a: order, P: placing
b: book, Q: delivering
Specification of f: $\forall x\ (P(x) \Rightarrow Q(f(x)))$

The weather forecast paradox

The weather forecast b is the product of an algorithm f (numerical analysis) to an input a (sensor data)
If f, a an explanation of $Q(b)$ (Q = "the temperature in Tokyo is", $b = 19°\mathrm{C}$)

No because these algorithms are expressed with huge programs that run for hours on massively parallel machines (nobody can trace the computation step by step)

Unlike the proof of

$$\forall n \in [1, 9]\ 12345679 \times 9 \times n = 111111111 \times n$$

that is a polynomial algorithm

So what makes an explanation is

- the width of its possible inputs (the generality of the statement)

- the small size of the program expressing the algorithm f
- the small execution time of f

Well-known constraint in ethics: the interlocutor must be able to understand the explanation

So we understand why

- the proof of the four color theorem is not an explanation (not generic, hence too big)
- the weather forecast is not explanation (too long)
- data-centric algorithms are not explanations (too big, when data is included)

From logic to ethics

Small, fast, wide algorithms as explanation fulfill the values of transparency, equality, agency, etc.