Explanation: from ethics to logic

Gilles Dowek

What is digital ethics?

The development of informatics (computer science, data science, digital technologies...) gives us an immense capacity

We can use this capacity for doing good or for doing evil
The same image recognition algorithms are used for medical imaging and mass surveillance

Ethics is the quest of an answer to the question How can I use this capacity for doing good?

A relatively recent concern

Compared to bio-ethics

The concepts of ethics

- Rules and consequences
  You ought to give a coin to a beggar because a (personal or collective) rule says so
  You ought to give a coin to a beggar because he will starve otherwise
  Discuss rules (from a code of conduct to a European law)
  Ask yourself who is going to use the piece of software you are developing
- Values
  Transparency, freedom of speech, benevolence, respect...
  Ask yourself what your values are and whether your actions (writing code, collecting data...) realize these values

The concepts of ethics

- Person (ethics of care)
  Ethics is not just about abstract concepts (rules, state of affairs, values) but also about people
  How to fight hate speech on social media? Understand that the abused people are people like you. Talk to them.
  Condemn the abusers to hug their victims.

Doing good is not always easy (1)

Moral dilemma: your values include : (1) freedom of speech and (2) the respect for victims of a crime against humanity
You moderating a web platform where some people glorify a genocide. Two options:

- let them do (and realize: freedom of speech)
- remove these messages (and realize: respect for the victims of a crime against humanity

No good way to act: find a compromise

Intercultural ethics: often (not always) people have the same values, but not the same value hierarchy
(1) Recognize you share the same values and you are both facing the same dilemma
(2) Do one step in the direction of the other

Doing good is not always easy (2)

A hospital decides to share all its medical files with a research lab

Sharing these files helps medical research
But it jeopardizes the patients' privacy

A solution: anonymize the files

But removing the last name of the patients is not enough
Wolfgang Amadeus *** (Salzburg, 1756 - Vienna, 1791), composer

How (and to which extent) can we anonymize a file? A research problem

An Ethical Turn in the Digital Industry?

What is a turn?

In the history of the digital industry

- Some companies completely mastered a technology (and reached a monopoly)
- But a new technology emerged
- They saw it, but could not adapt, and lost their leadership to another company

Examples:
- IBM mastered hardware technology, but could not adapt to the emergence of software
- Microsoft mastered software technology (os...), but could not adapt to the emergence of the Web
- Google mastered (asymmetric, one-to-many) Web, but could not adapt to the emergence of the symmetric Web (social media...)
- ...

What is next?

Difficult to tell

But we can answer a simpler question: what is wrong in the current situation? (and what makes users unhappy?)

- A social media business model based of the selling of the user's privacy
- hate speech
- on-line accusation without proof (revenge rather than justice)
- fake news

All these are ethical questions: privacy, peace, justice, truth

May be (remember we are not prophets)

- the emerging technology is ethics
- the current leaders will not manage to adapt (empirical fact)
- and new leaders respecting privacy, peace, justice, and truth will emerge

Ethics and innovation

Ethics restricts innovation

If ethics is the next turn then the exact opposite:
Ethics is a necessary condition to innovation

Explanation: from ethics to logic

Explanation: an ethical necessity

My bank loan application has been rejected
I want an explanation

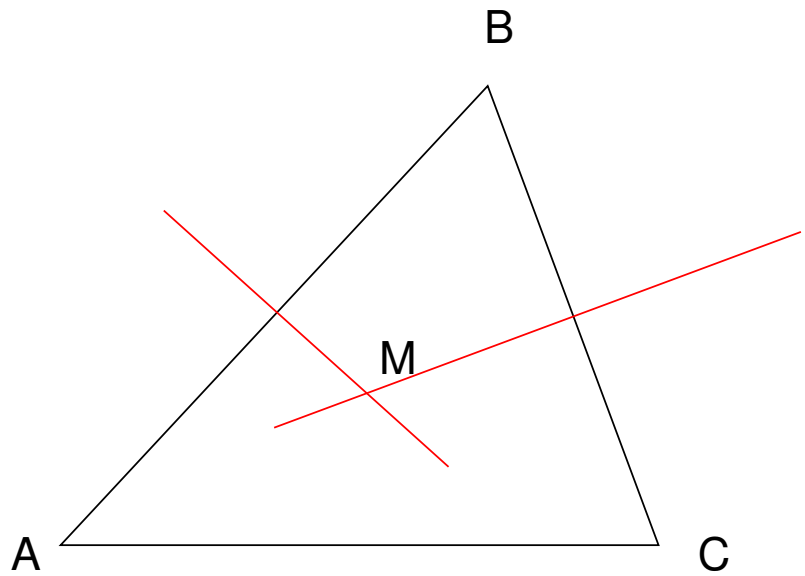
Transparency: I want to know the rules
Equality: I want to be sure the same rules apply to everyone
Agency: I want to be able to "improve" and apply again
Dignity: I want to be considered as a rational being

Explanation itself as a value?

Proof as an explanation

What is explained? A statement

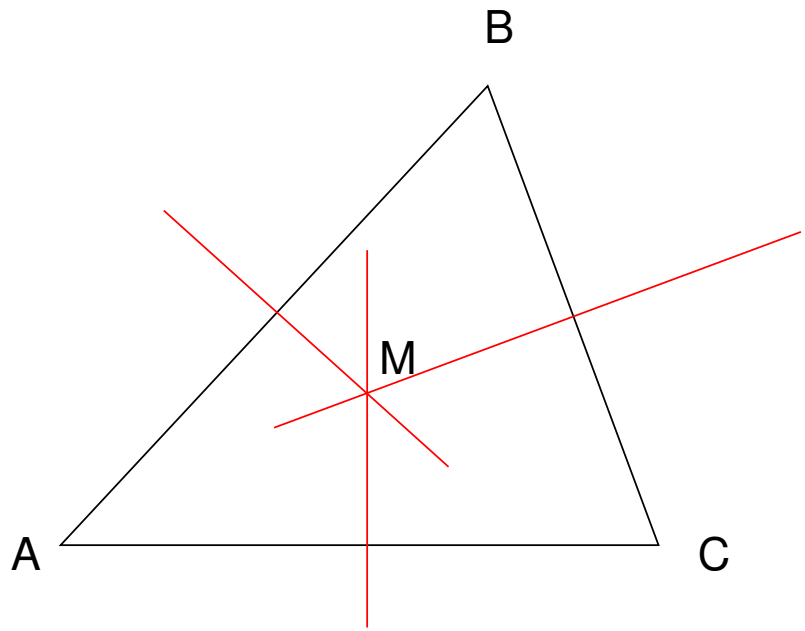
What explains it? A proof of this statement



Proof as an explanation

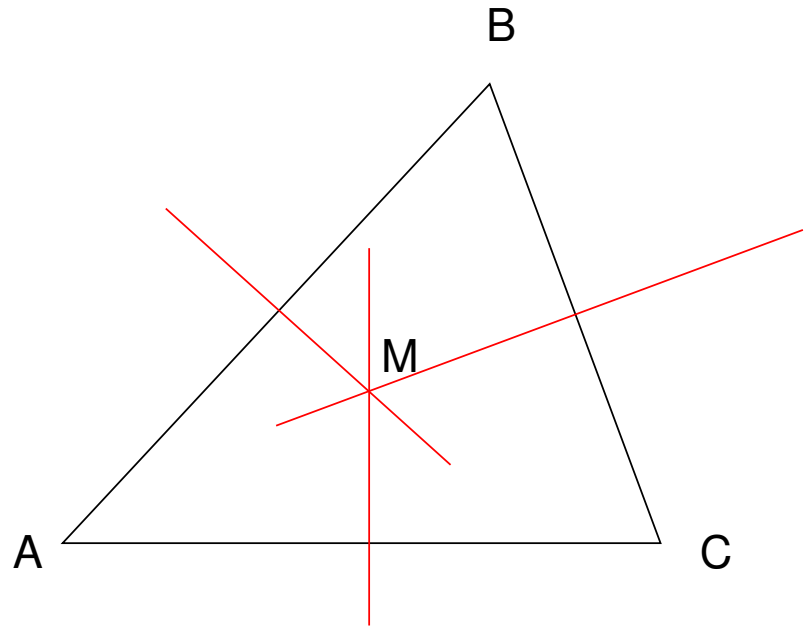
What is explained? A statement

What explains it? A proof of this statement



Proof as an explanation

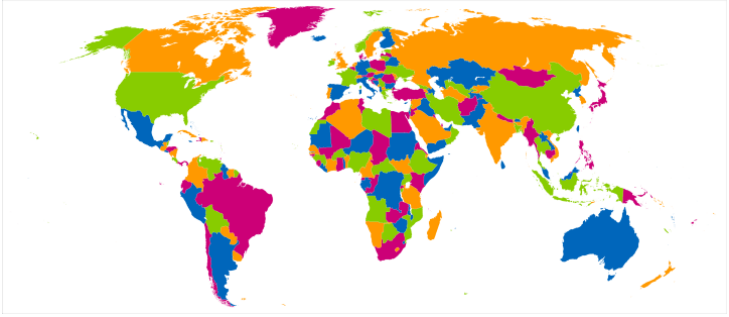
What is explained? A statement
What explains it? A proof of this statement



$d(M, A) = d(M, B)$ 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

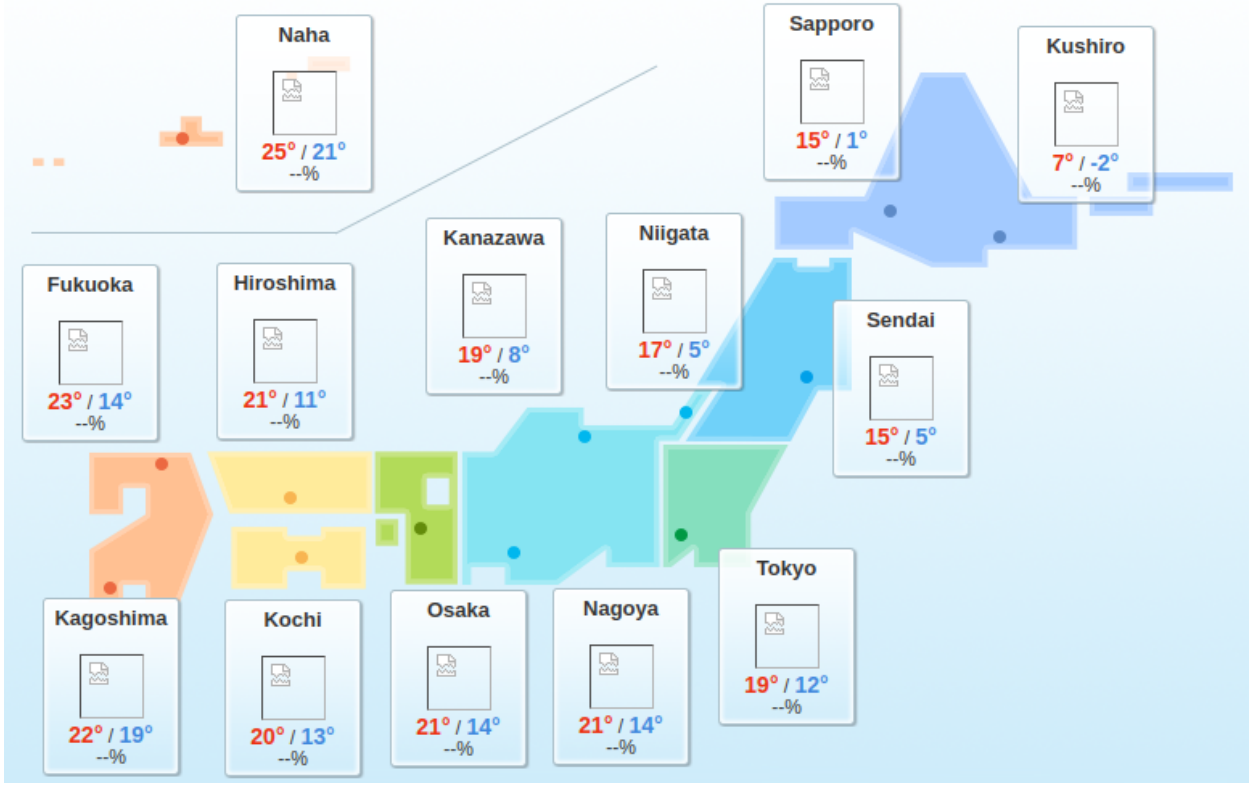


Every map is four colorable
Reduce the infinitude of possible maps to 1482
Check that these 1482 maps are "reducible": case study
"First" proof by computer

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

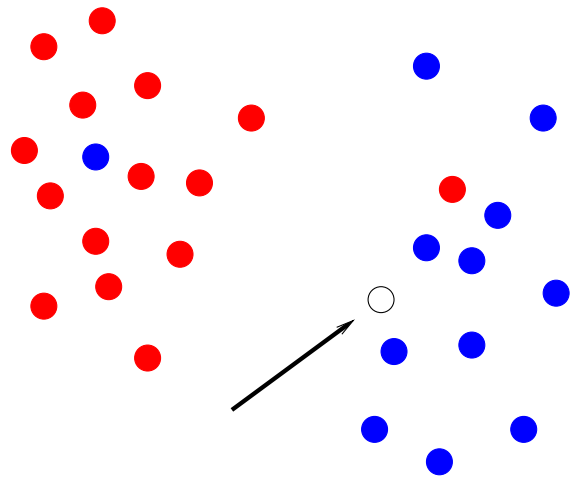
Weather forecast: a very long numerical analysis computation leading to the map



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

And sometimes difficult to tell what an explanation would look like

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

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

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

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

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

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

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

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

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

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

$$\begin{array}{r} 12345679 \\ \times\quad 36 \\ \hline 74074074 \\ 37 \end{array}$$

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

$$\begin{array}{r} 12345679 \\ \times\quad 36 \\ \hline 74074074 \\ 70370 \end{array}$$

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

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

$$
\begin{array}{r}
12345679 \\
\qquad\quad 36 \\
\hline
740740740 \\
703703703
\end{array}
$$

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

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

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

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

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

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

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

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

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

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

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

$$12345679 \times 36 = 444444444$$

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)

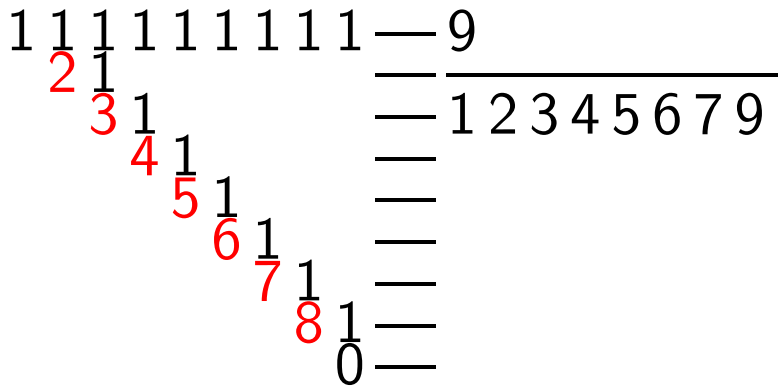
$$\forall n \in [1, 9]\ 12345679 \times 9 \times n = 111111111 \times n$$

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 $111111111/9 = 12345679$



The n-th digit of the result is n, also the n-th partial remainder is $n + 1$

By induction: the $(n + 1)$-th partial dividend is

$10(n + 1) + 1 = 9(n + 1) + n + 2$

$(n + 1)$-th digit of the result: $n + 1$, $(n + 1)$-th partial remainder: $n + 2$

Works only when $n + 2 < 9$, ~~the eighth digit is 8 and the eighth partial remainder is 9~~ the eighth digit is 9 and the eighth 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 the statement $12345679 \times 36 = 444444444$ 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 deducing 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

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

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, unless we establish that all images with a black pixel at position (314,42) are recognized as dogs (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

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 that maps every number p in $[1,9]$ to a proof of the statement
$12345679 \times 9 \times p = 111111111 \times p$

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 (Kolmogorov)
- the small execution time of f (Bennett)

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.