Does explaining AI make it safer?

Husky vs Wolf:

had the background been green, this would be a wolf.

AI software is safe if it does not cause any direct harm and cannot be misused by adversaries to do so. Some AI techniques cannot be verified due to their complexity, however an explanation may be a good proxy -- **counterfactuals** are a go-to method.

Kacper Sokol and Peter Flach, University of Bristol, UK

Counterfactual explanations -- what can go wrong.

An explanation helps to:

- **understand**, **monitor** and **verify** (**debug**) behaviour of a system (husky vs wolf) -- comprehensible by domain experts and a lay audience alike;
- assess **fairness** of a system (...had this person been a female, the outcome...).

However, they are:

- **local** and may **not generalise** (does the green background affect all dog images?);
- **observational** and **not causal** (does white background always indicate a husky?);
- very similar to **adversarial examples** but the *perturbation is meaningful*:
  - using the insights extracted from an explanation to **game a model** -- (injecting green background),
  - counterfactuals **leak information** about the *model's decision logic* and the *training data*.