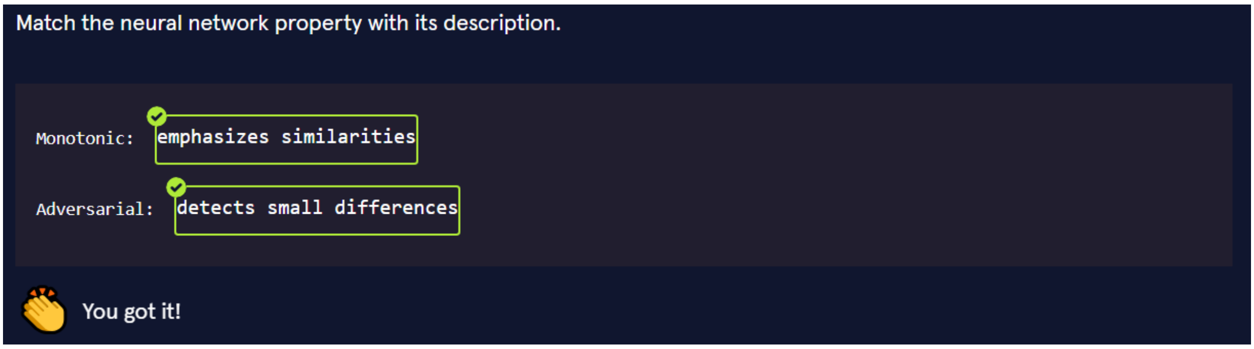
QUIZ

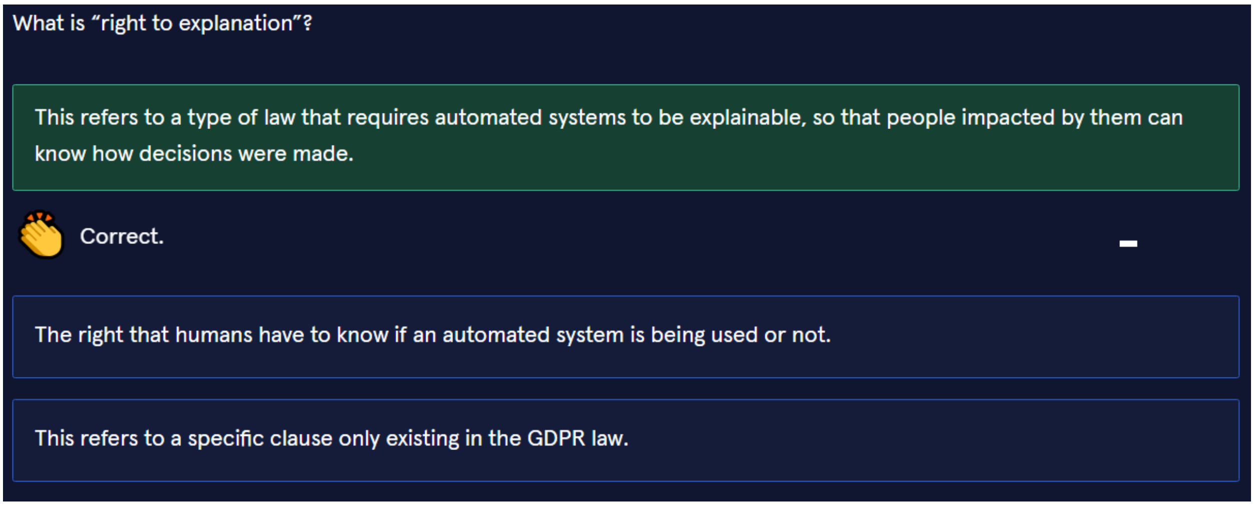


Match the neural network property with its description.

Monotonic: ✓ emphasizes similarities

Adversarial: ✓ detects small differences

👏 You got it!



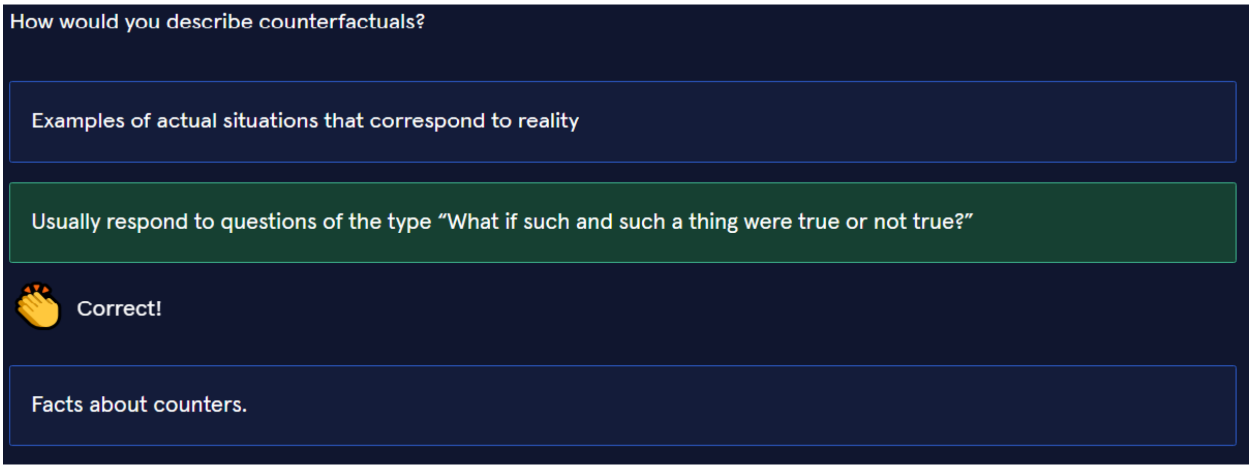
What is “right to explanation”?

This refers to a type of law that requires automated systems to be explainable, so that people impacted by them can know how decisions were made.

👏 Correct.

The right that humans have to know if an automated system is being used or not.

This refers to a specific clause only existing in the GDPR law.



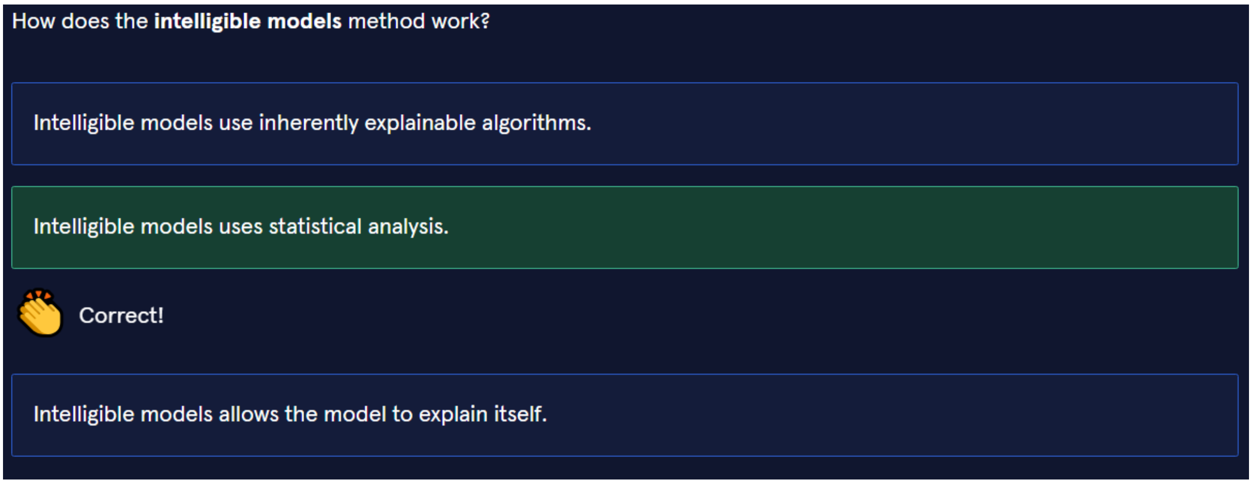
How would you describe counterfactuals?

Examples of actual situations that correspond to reality

Usually respond to questions of the type “What if such and such a thing were true or not true?”

👏 Correct!

Facts about counters.



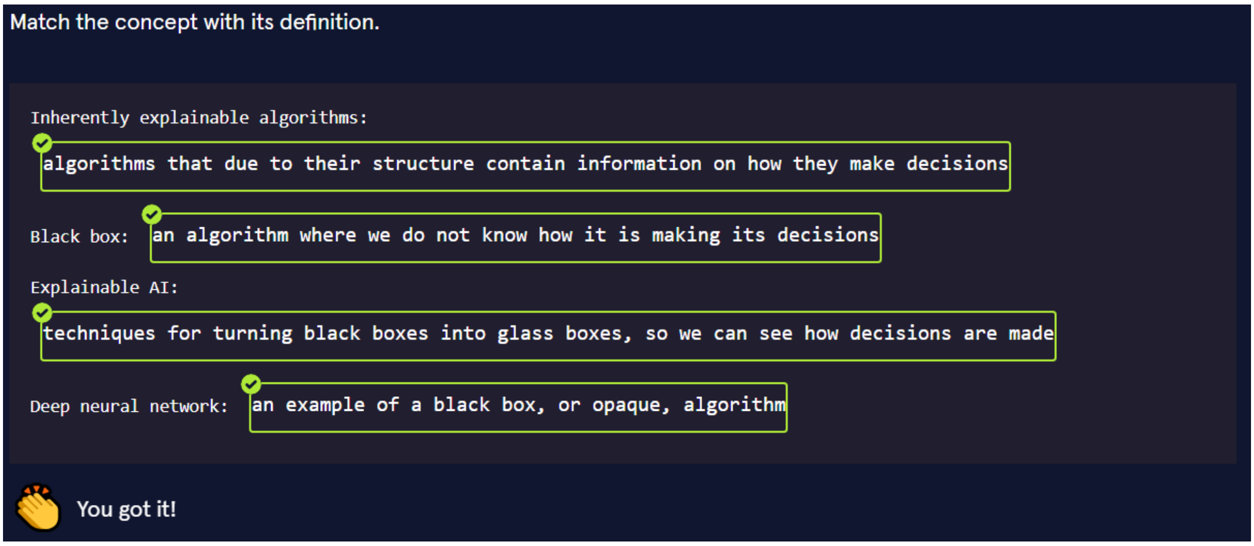
How does the **intelligible models** method work?

Intelligible models use inherently explainable algorithms.

Intelligible models uses statistical analysis.

👏 Correct!

Intelligible models allows the model to explain itself.



Match the concept with its definition.

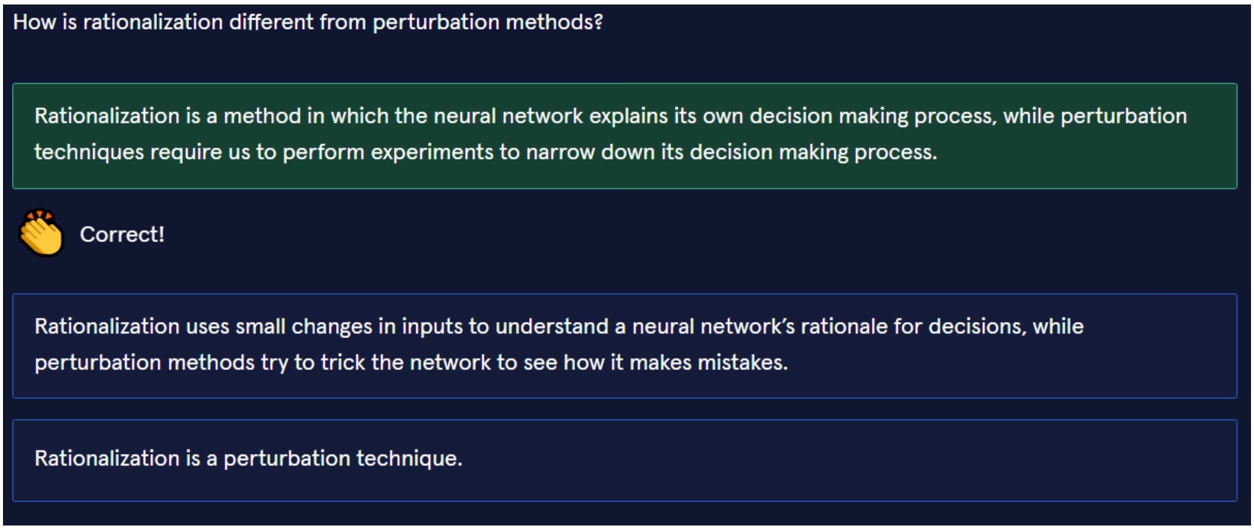
Inherently explainable algorithms:
✓ algorithms that due to their structure contain information on how they make decisions

Black box: ✓ an algorithm where we do not know how it is making its decisions

Explainable AI:
✓ techniques for turning black boxes into glass boxes, so we can see how decisions are made

Deep neural network: ✓ an example of a black box, or opaque, algorithm

👏 You got it!



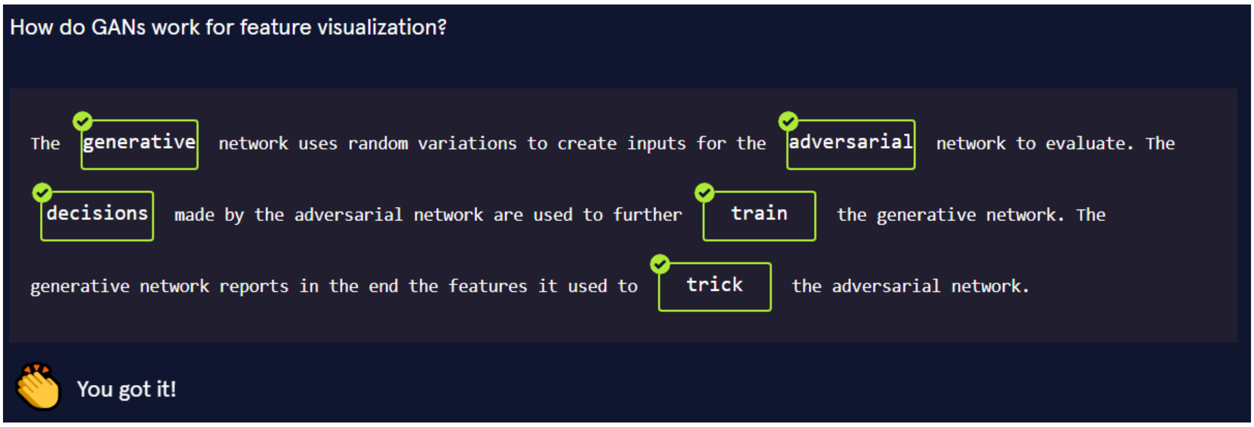
How is rationalization different from perturbation methods?

Rationalization is a method in which the neural network explains its own decision making process, while perturbation techniques require us to perform experiments to narrow down its decision making process.

👏 Correct!

Rationalization uses small changes in inputs to understand a neural network’s rationale for decisions, while perturbation methods try to trick the network to see how it makes mistakes.

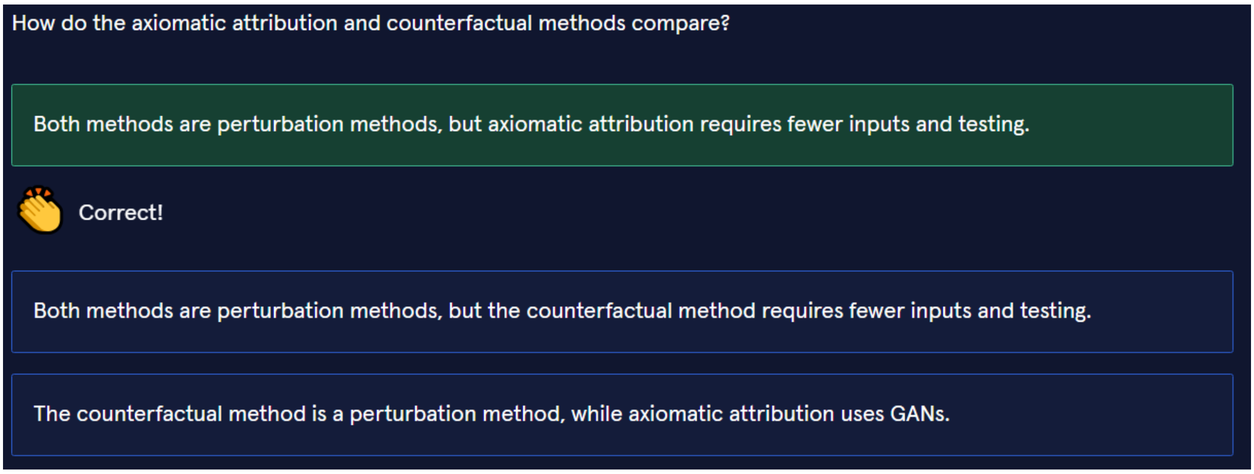
Rationalization is a perturbation technique.



How do GANs work for feature visualization?

The ✓ generative network uses random variations to create inputs for the ✓ adversarial network to evaluate. The ✓ decisions made by the adversarial network are used to further ✓ train the generative network. The generative network reports in the end the features it used to ✓ trick the adversarial network.

👏 You got it!



How do the axiomatic attribution and counterfactual methods compare?

Both methods are perturbation methods, but axiomatic attribution requires fewer inputs and testing.

👏 Correct!

Both methods are perturbation methods, but the counterfactual method requires fewer inputs and testing.

The counterfactual method is a perturbation method, while axiomatic attribution uses GANs.