Counterfactual Explanations

MAKING AI DECISION-MAKING MORE USEFUL AND TRUSTWORTHY

Cross-industry | Financial Services | Healthcare | Health & Public Service | Speciality chemicals | FMCG |

The Alan Turing Institute

accenture

Innovation Symposium 2020

Counterfactual Explanations

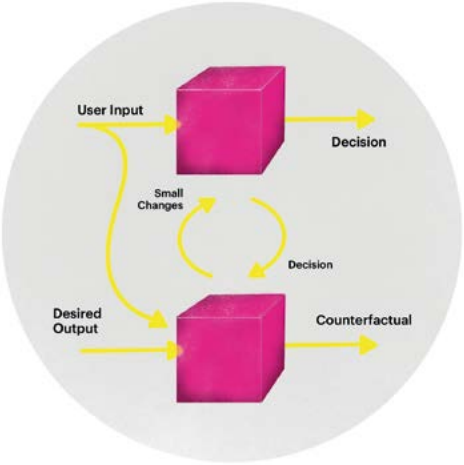
Want a sneak preview of the research? Click to hear from our presenter.

MAKING AI DECISION-MAKING MORE TRUSTWORTHY AND USEFUL

The challenge

The inner workings of AI models can be complex and hidden, particularly with black box models. This means it's hard to explain how AI systems reach their decisions—like declining a loan, for example.

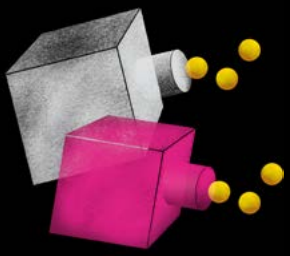
Counterfactual Explanations help to uncover these decisions, and to identify changes in conditions that would result in the desired outcome —like how much salary should increase, or the decrease in debt needed. How can we ensure these explanations are useful and trustworthy from a human point of view?



We can validate counterfactuals based on human input, and by assigning a confidence score.

Our research

Last year, we created a proof of concept for counterfactual explanations. We piloted with a company in the FMCG space. Building on this knowledge, we're now working to make counterfactuals more robust and actionable. Our latest research aims to evaluate the confidence of the counterfactual explanation produced. We're also working to incorporate human input when optimising the counterfactual model, to create more useful, trustworthy explanations.



To find out more, contact Lisa Schut:
lisaschut94@gmail.com

Project team:
Lisa Schut, Research Assistant, OATML Group, The University of Oxford; Yarin Gal, The Alan Turing Institute; Rory McGrath, Bogdan Sacaleanu, Accenture.