# Counterfactual Explanations and Algorithmic Recourse

Patrick Altmeyer (student)
Dr. Cynthia Liem (supervisor)

TU Delft Delft University of Technology

## Explaining black box models through counterfactuals

Human operators in charge of the black-box decision-making systems do not understand how they works and essentially often they still choose to rely on it blindly. Unfortunately, those individuals who are subject to the decisions produced by such systems typically have no way of challenging them. **Counterfactual Explanations (CE) can help programmers make sense of the systems they build**: they explain how inputs into a system would have to change for it to produce a different output. CEs that involve realistic and actionable changes can be used for the purpose individual recourse: **Algorithmic Recourse (AR) offers individuals subject to algorithms a way to turn a negative decision into positive one.**

Our work so far:

- Built a scalable library for **C**ounterfactua**L** **E**xplanations and **A**lgorithmic **R**ecourse in Julia: CLEAR.jl.
- Have run experiments investigating the dynamics of AR and proposed a related research project to bachelor's students.

## From basic principles …

### A light-hearted motivating example

Suppose we have fitted some black box classifier to divide cats and dogs based on two features: height and tail length. One individual cat – let's call her Kitty 🐱 – is friends with a lot of cool dogs and wants to remain part of that group. The counterfactual path below shows how 🐱 needs to change her appearance in order to be allocated to the group of dogs by the system.

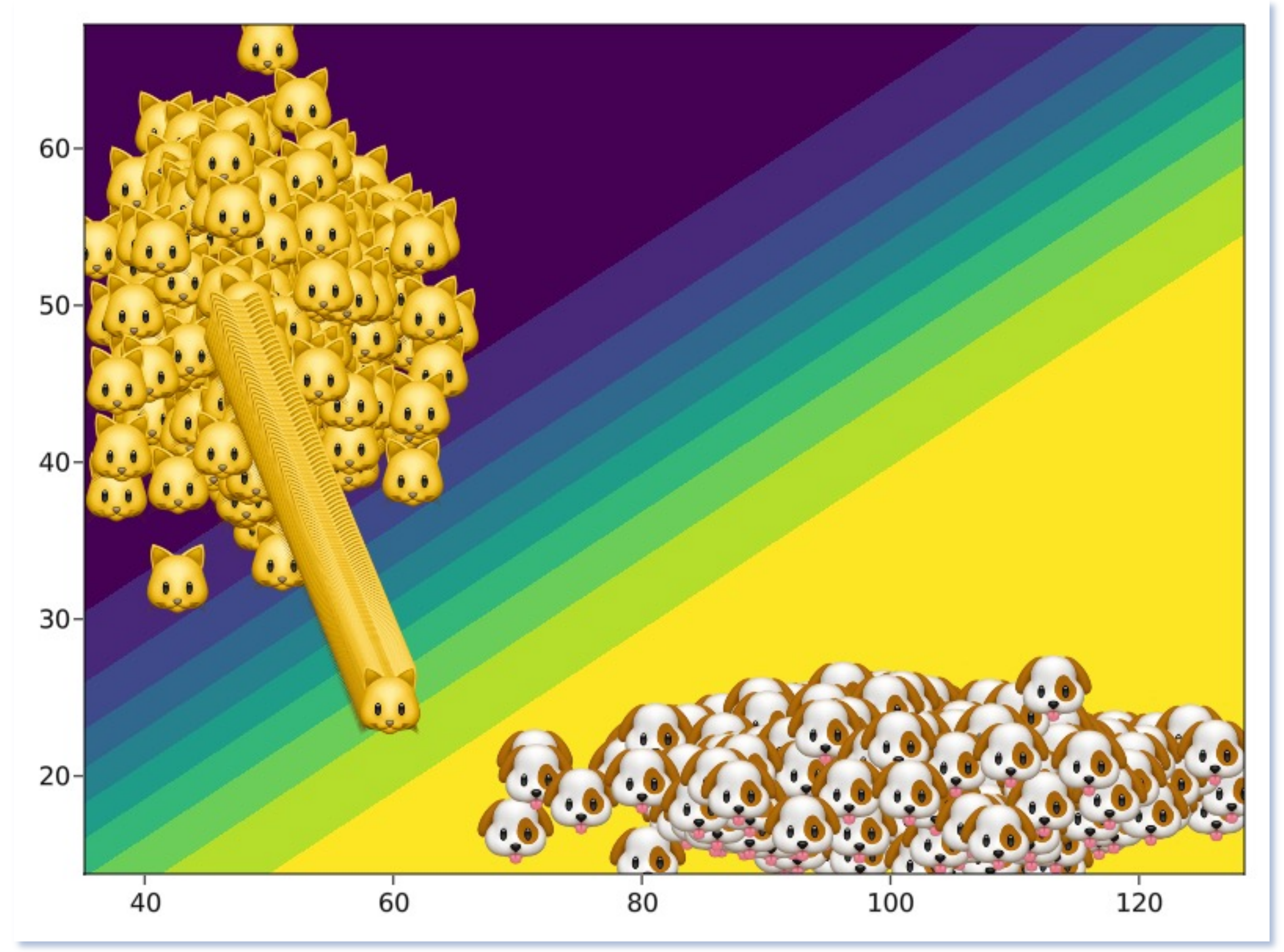

Figure 1: Generating recourse for 🐱 following Wachter et al. (2018). Contour shows the predictions of a simple MLP.

## … to realistic recourse.

### CE through minimizing predictive uncertainty

As 🐱 crosses the decision boundary she fools the system, but we can still clearly distinguish her from the rest of her dog friends. Her counterfactual self is **ambiguous** and **unrealistic**.
Consider instead the counterfactual path generated in Figure 2: for the same confidence threshold, 🐱 ends up in much denser area.

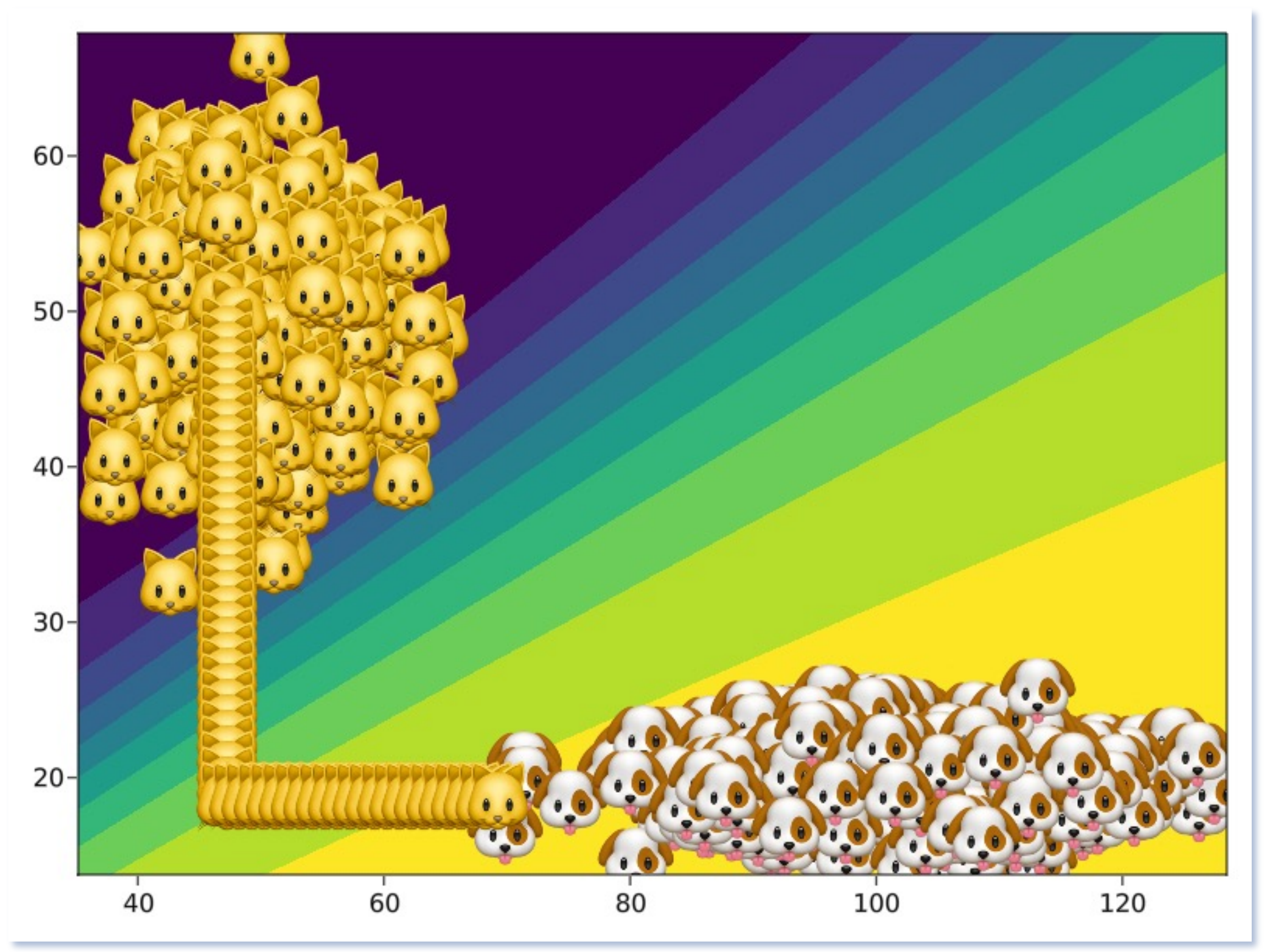

Figure 2: Generating recourse for 🐱 following Schut et al. (2021). Contour shows the predictions of a Bayesian MLP.

The method used to generate the counterfactual in Figure 2 is fast and greedy approach that works by minimizing the predictive uncertainty of Bayesian models. Applied to MNIST data the Bayesian approach generates satisfactory outcomes (Figure 3).

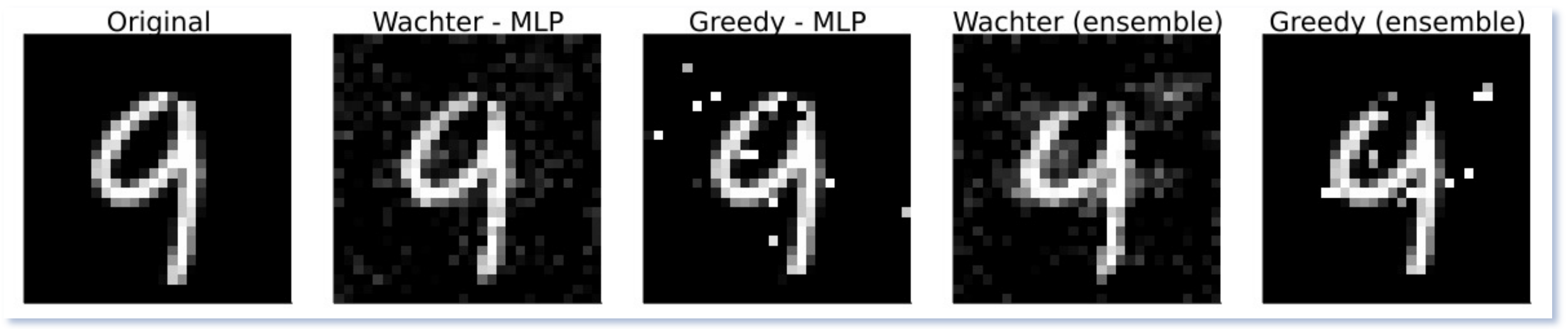

Figure 3: Turning a 9 into a 4: counterfactual explanations for MNIST data. For the MLP counterfactuals look like adversarial attacks. Counterfactuals for the deep ensemble are arguably much better.

## But wait a minute …

### Beyond the static setting

In practice decision-making systems are regularly updated. Recent work has investigated the robustness of AR: can we be sure that 🐱 can stay with her dog friends after model updates? In our work we go a step further and ask ourselves: does 🐱 herself trigger model shifts through her move across the decision boundary? Does that have consequences for other cats or dogs that want to implement recourse? More generally, **what are the dynamics of algorithmic recourse?**

## The dynamics of Algorithmic Recourse

### Endogenous domain and model shifts

Our preliminary experiments show that even for Bayesian models AR essentially generates new clusters of individuals. While these clusters remain on the target side of the decision boundary, they could still be distinguished from individuals that were always in the target class. This may leave them vulnerable to discrimination through the system.

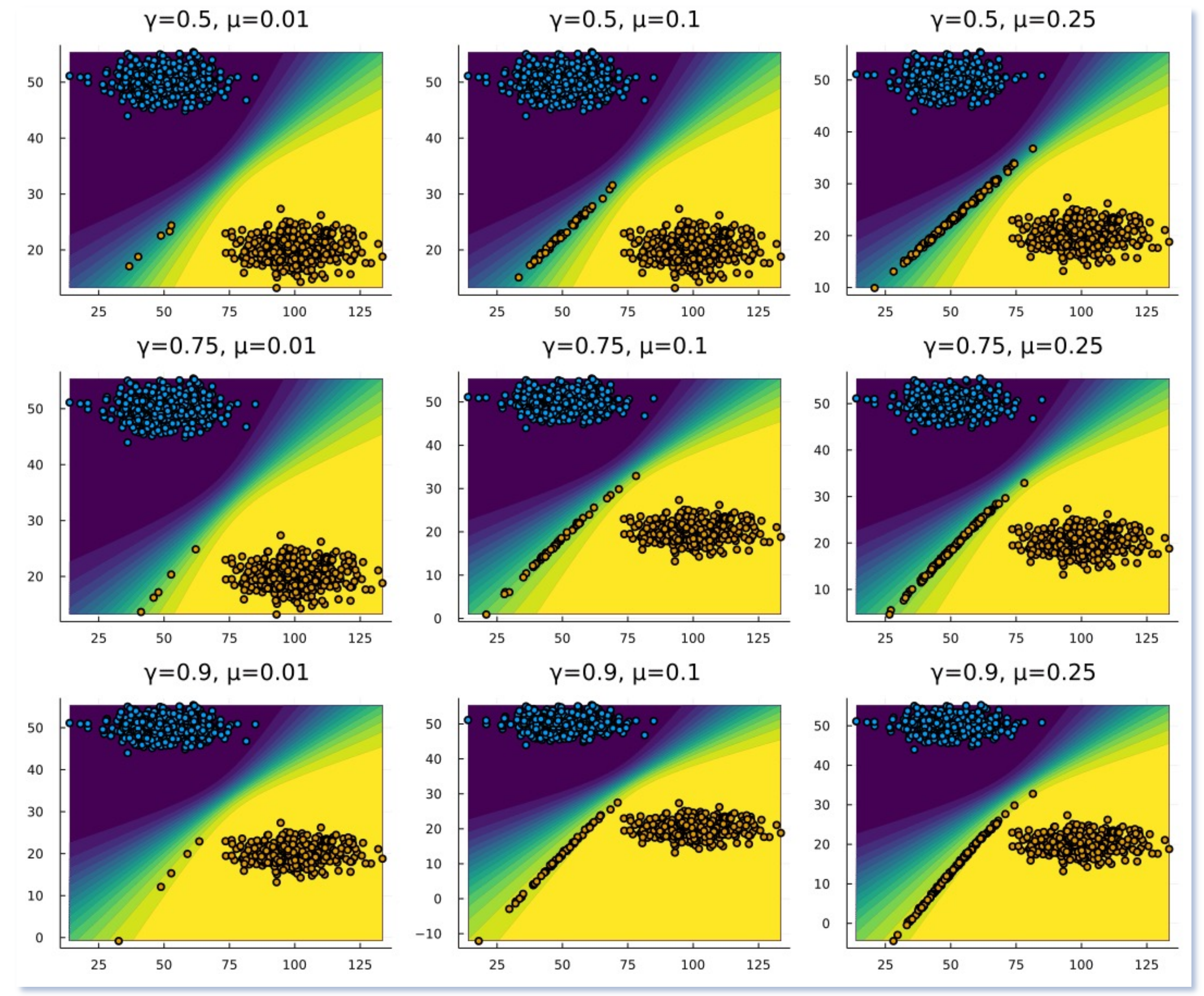

Figure 4: Algorithmic recourse leads to domain and model shifts.

## Where to go from here

### Open questions (your thoughts are more than welcome!)

- How much of an issue is this really? Can we think of real-world examples where scope for discrimination may lead to undesirable outcomes?
- How does the magnitude of domain and model shifts vary across different approaches to generating AR? (student project)
- Can we assess what factors mitigate endogenous shifts when generating recourse?

### References

Wachter et al. (2018). "Counterfactual explanations without opening the black box: automated decisions and the GDPR.". In: Harvard Journal of Law & Technology (31)

Schut et al. (2021). "Generating interpretable counterfactual explanations by implicit minimisation of epistemic and aleoteric uncertainty.". In: Proceedings of Machine Learning Research (130)