

What's new in Trustworthy AI in JuliA?

JuliaCon 2024

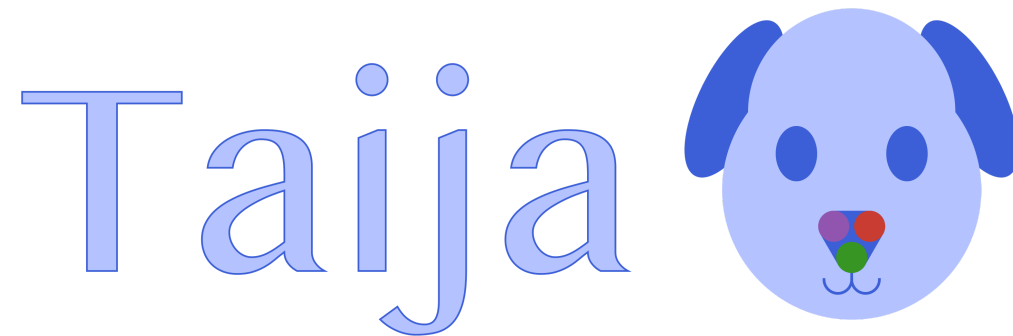
Patrick Altmeyer

Thursday, July 11, 2024



What's Taija?

Taija is a small but growing ecosystem of packages geared towards **T**rustworthy **A**I in **J**uli**A**—**T**aija



Trustworthy AI in Julia: github.com/JuliaTrustworthyAI

Ecosystem Overview

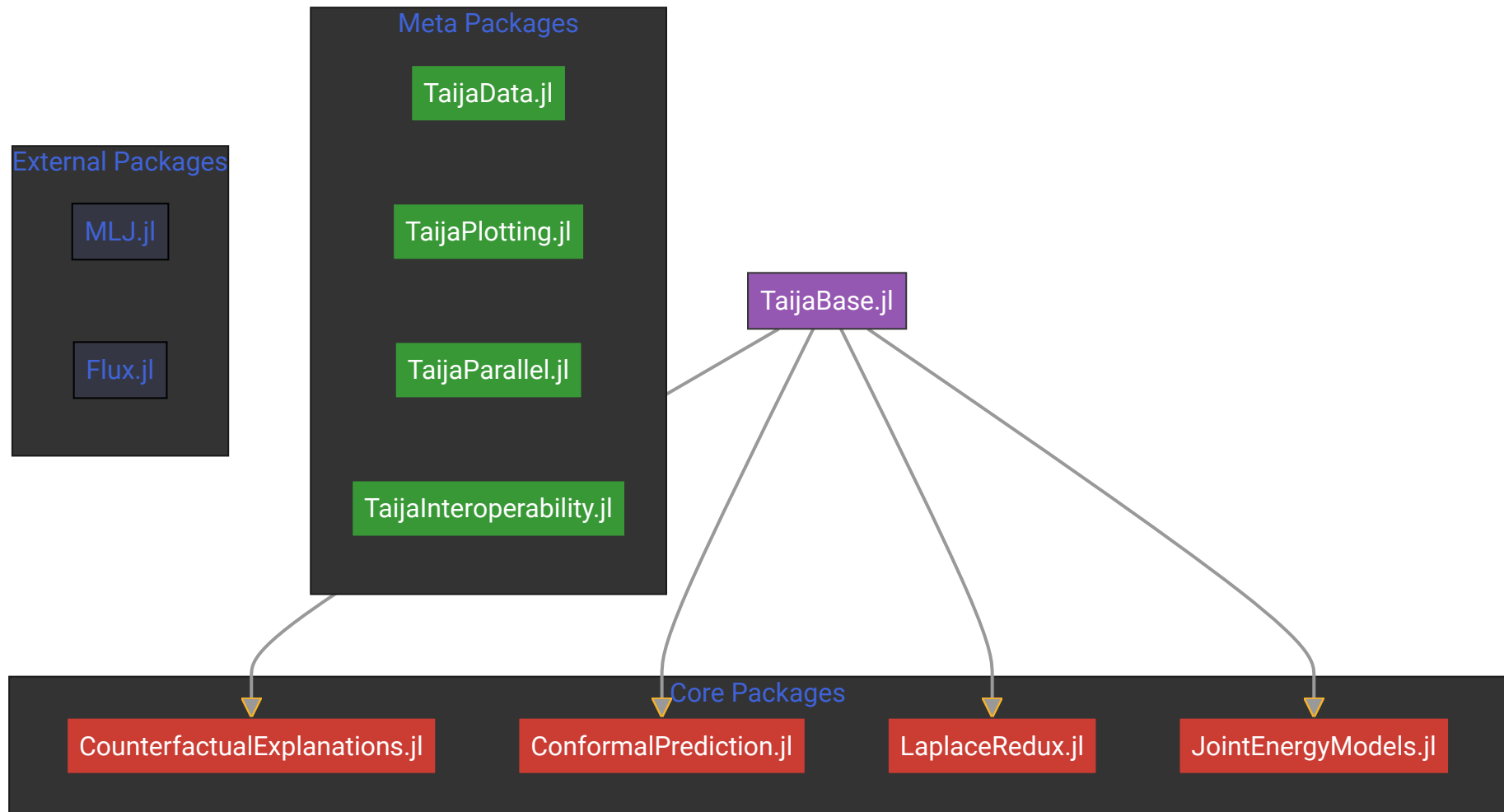


Figure 1: Overview of the Taija ecosystem. Early-stage packages omitted.


Who's behind Taija?

- **2021-2022:** Initially developed by myself to support my PhD research in Trustworthy AI (still ongoing).
- **2023-2024:** Expanded by a growing community of contributors, including TU Delft and G/JSoC students.

Thanks to @MojiFarmanbar, @JorgeLuizFranco, @Rockdeldiablo, @kmariuszk, @RaunoArike, @VincentPikand, @severinbratus, @rithik83, @navimakarov, @laurikskl, @MarkArdman, @adelinacazacu, @Andrei32Ionescu and many others!

Use Cases

Who could benefit from Taija?

- Researchers in AI and ML, particularly in the fields of explainability, uncertainty quantification, and Bayesian deep learning: (, ...)
- Practitioners using conventional ML and DL models who are interested in understanding the models' decisions and their uncertainty.
- Julia developers who want to contribute to the ecosystem (any level of expertise is welcome!).

Research

Counterfactual Explanations

The largest single category of CE methods solves the following optimization through gradient descent:

$$\mathbf{s}^* = \arg \min_{\mathbf{s}' \in \mathcal{S}} \{ \text{yloss}(M(f(\mathbf{s}')), y^*) + \lambda \text{cost}(f(\mathbf{s}')) \}$$

Pick your Poison

All of these counterfactuals are valid explanations ...

... which one would you pick?

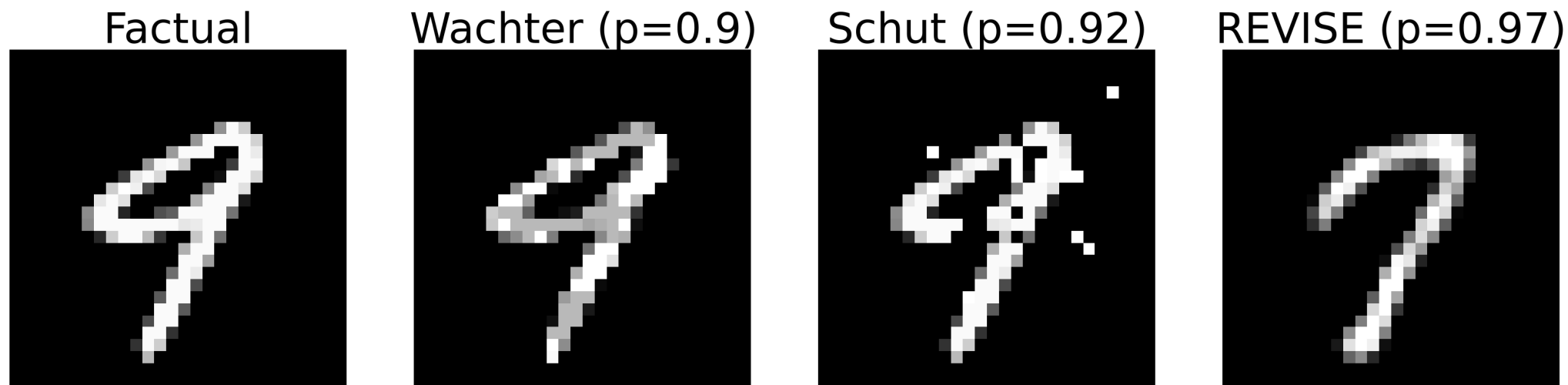


Figure 3: Turning a 9 into a 7: Counterfactual explanations for an image classifier produced using *Wachter* (Wachter, Mittelstadt, and Russell 2017), *Schut* (Schut et al. 2021) and *REVISE* (Joshi et al. 2019).

Faithful Counterfactuals

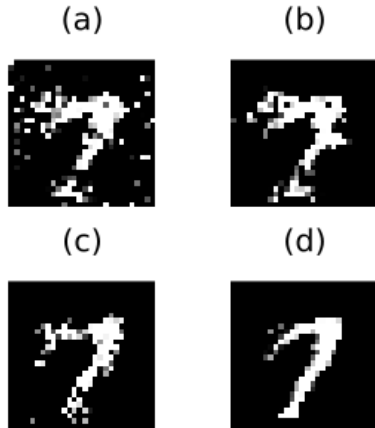


Figure 4: Turning a 9 into a 7. *ECCTCo* applied to MLP (a), Ensemble (b), JEM (c), JEM Ensemble (d).

*ECCTCo*¹ counterfactuals

- explain models faithfully (Figure 4).
- achieve SOTA plausibility (Figure 5).

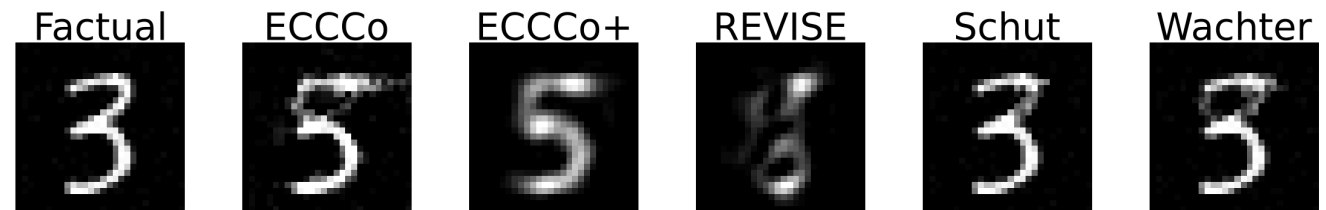


Figure 5: Results for different generators (from 3 to 5).

1. [Package](#) and link to AAAI 2024 paper: github.com/pat-alt/ECCTCo.jl.

Intent Classification

Intent classification (IC) in dialogue systems is a common task and a natural place for conformal prediction.

- Simply returning top-1 softmax likely wrong.
- Existing ad-hoc approach is top- k .
- Conformal classifiers predict sets that fulfill coverage guarantee.

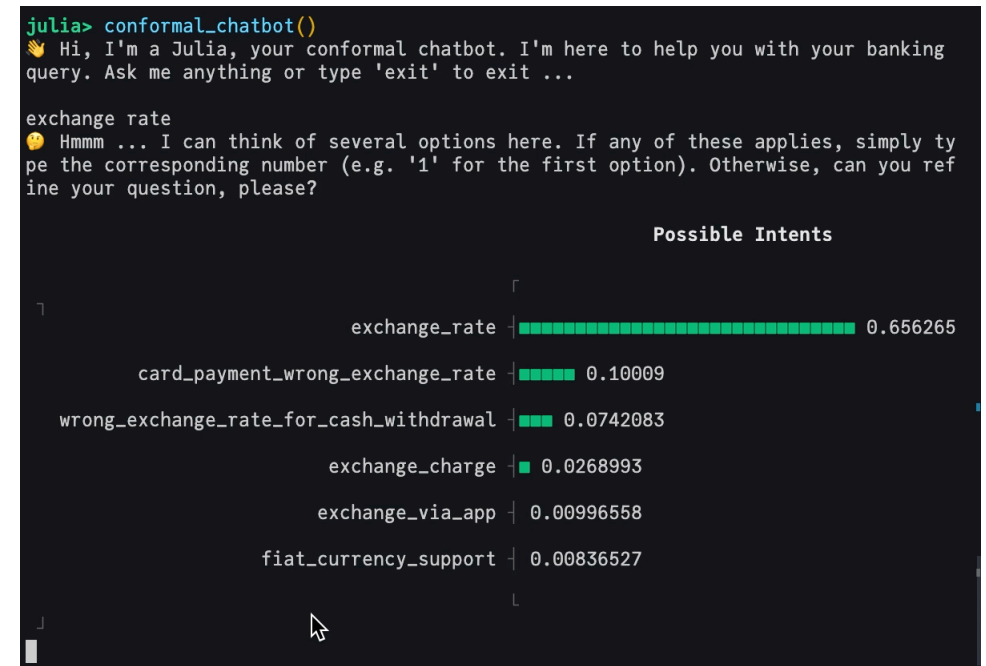


Figure 6: A simple conformal chat bot in the Julia REPL using ConformalPrediction.jl.

Conformal IC and Clarification

- Our NAACL 2024 [paper](#) introduces CICC: a framework for fast and accurate intent classification in conversational AI.
- Winning project at ING Experiment Week 2023.

See also *Building a Conformal Chatbot in Julia* with `ConformalPrediction.jl` and `Transformers.jl`¹.

1. Experiments in Hengst et al. ([2024](#)) were run in parallel using Python's MAPIE and `ConformalPrediction.jl`, in order to cross-check results. Reported results were produced using MAPIE.

More Research

- *Stop Making Unscientific AGI Performance Claims* ([Altmeyer et al. 2024](#)) [upcoming](#) at *ICML 2024*.¹
- *Endogenous Macrodynamics in Algorithmic Recourse* ([Altmeyer et al. 2023](#)) [published](#) at *IEEE SaTML 2023*.
- Various Master's theses on CE for imbalanced data ([Zagorac 2024](#)), CE for LLMs ([draft PR](#)), and more ...
- Bachelor's theses on *What Makes Models Explainable? Evidence from Counterfactuals* (related development: [AdversarialRobustness.jl](#)).

1. Our related [package](#) is not currently part of Taija but may be in the future.

Developments

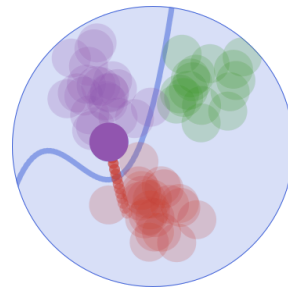
A brief overview of the highlights. For a more introductory presentation, see also this [slides](#) presentation The Alan Turing Institute



What's new in ...

`CounterfactualExplanations.jl`: A [package](#) for Counterfactual Explanations and Algorithmic Recourse in Julia.

- refactoring (e.g. package extensions)
- performance improvements
- new features



Counterfactual
Explanations

- JCon Proceedings ([Altmeyer, Deursen, et al. 2023](#)).

Composable Generators

Recall that for most generators, we have:

$$\mathbf{s}^* = \arg \min_{\mathbf{s}' \in \mathcal{S}} \{ \text{yloss}(M(f(\mathbf{s}')), y^*) + \lambda \text{cost}(f(\mathbf{s}')) \}$$

Why not compose generators that combine ideas from different off-the-shelf generators?

```
1 @chain generator begin
2     @objective logitcrossentropy
3     + 1.0ddp_diversity      # DiCE (Mothilal et al. 2020)
4     @with_optimiser Flux.Adam(0.1)
5     @search_latent_space    # REVISE (Joshi et al. 2019)
6 end
```

Explaining Different Models

Besides any Flux.jl model, extensions add support for

- [DecisionTree.jl](#)
- [NeuroTrees.jl](#) (see Jeremie's [talk](#) Fri 10:10–10:20 For Loop)
- [LaplaceRedux.jl](#)
- [JointEnergyModel.jl](#) (upcoming)

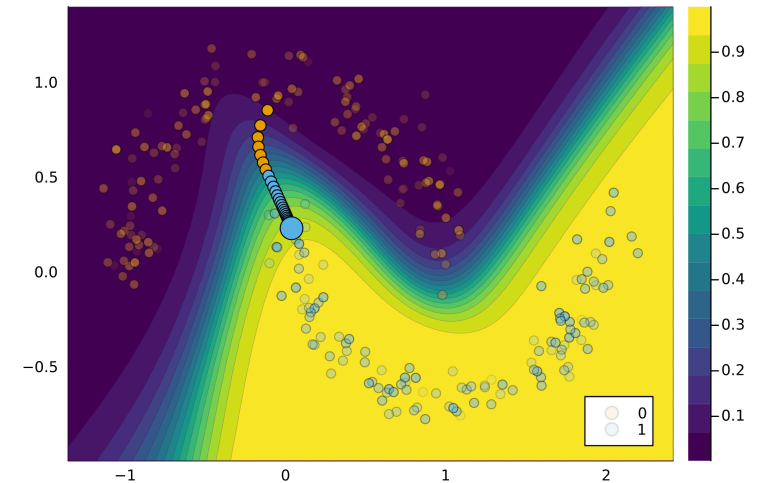


Figure 7: Counterfactual for differentiable decision tree classifier.

Benchmarking Explanations

Extensive support for evaluating and benchmarking explanations.

Evaluation

```
1 # Generate counterfactuals
2 ces = generate_counterfactuals(
3     factual,
4     target_label,
5     data,
6     M,
7     generator;
8     num_counterfactuals=5
9 )
10
11 # Evaluate them
12 evaluate(ces)
```

Benchmarks

Benchmark all available generators and models at once in parallel¹:

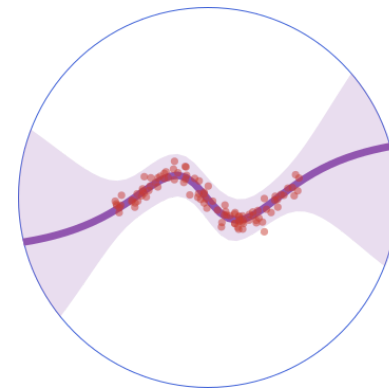
```
1 using TaijaParallel
2 pll = ThreadsParallelizer()
3 bm = benchmark(
4     counterfactual_data;
5     parallelizer = pll
6 )
```

1. [TaijaParallel.jl](#) adds support for parallelization. [Join](#) Friday 11:50–12:00, Else (1.3)

What's new in ...

`LaplaceRedux.jl`: A [package](#) for Effortless Bayesian Deep Learning through Laplace Approximation for Flux.jl neural networks.

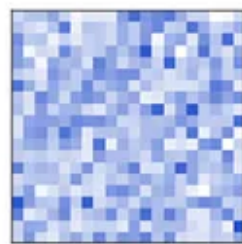
- new features
- interface to MLJ
- JCon Proceedings (under review).



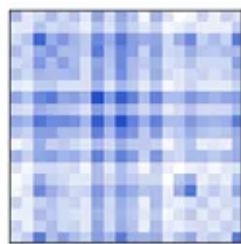
Laplace
Redux

Student Contributions

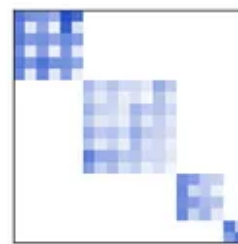
- Support for multi-class problems.
- Support for more scalable Hessian approximations.
- Interface to MLJ for easy model training and evaluation.



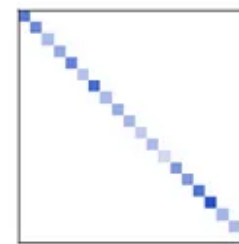
(a) Full



(b) LRank



(c) KFAC



(d) Diag.

Figure 8: Hessian approximations. Source: Daxberger et al. ([2021](#))

Check out their [blog post](#)!

What's new in ...

`ConformalPrediction.jl`: A [package](#) for Predictive Uncertainty Quantification through Conformal Prediction for Machine Learning models trained in MLJ.

- refactoring
- new features ([@MojiFarmanbar](#))
 - Time Series
 - Quantile Regression



Joint Energy Models

`JointEnergyModels.jl`: A package for Joint Energy-Based Models Models in Julia (early development).

Hybrid models that can predict and generate ([Grathwohl et al. 2020](#)).

- Flux.jl interface
- MLJFlux.jl interface

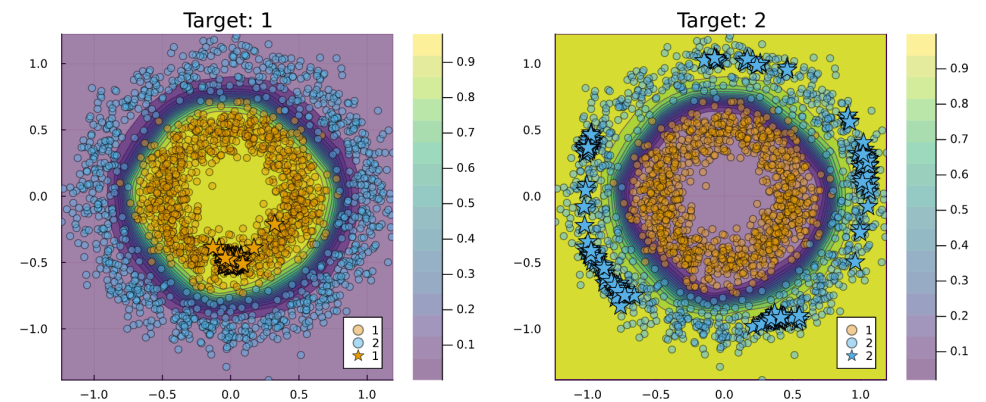


Figure 9: Predicted output class probabilities (contour) and generated inputs (stars) for a binary JEM classifier.

Final Things

Ongoing Work

Taija has been running two Julia Season of Code projects this summer.

1. (*Conformal Bayes*) Bridging the gap between Bayesian and frequentist approaches to Predictive Uncertainty Quantification with [@Rockdeldiablo](#) and co-supervisor [@MojiFarmanbar](#)
2. (*Causal Recourse*) From minimal perturbations to minimal interventions for Algorithmic Recourse with [@JorgeLuizFranco](#) and co-supervisor [@mschauer](#).

Student Testimonials

Students have generally been enthusiastic about their experience with Julia and Taija:

“Programming in Julia has definitely helped us become better programmers. [...] whenever we had such questions and asked them [to] the wider Julia community, there were always people ready to help in my experience, which was nice.”

— @RaunoArike

Get Involved

- Working on related projects?
- Interested in contributing to Taija?
- Want to learn more about Trustworthy AI in Julia?
- Any suggestions or feedback?

Get in touch with me or any of the contributors! Join our #taija channel on the JuliaLang Slack or visit our GitHub organization.

Questions?



References

- Altmeyer, Patrick, Giovan Angela, Aleksander Buszydlik, Karol Dobiczek, Arie van Deursen, and Cynthia CS Liem. 2023. "Endogenous Macrodynamics in Algorithmic Recourse." In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 418–31. IEEE.
- Altmeyer, Patrick, Andrew M. Demetriou, Antony Bartlett, and Cynthia C. S. Liem. 2024. "Position Paper: Against Spurious Sparks-Dovelating Inflated AI Claims." <https://arxiv.org/abs/2402.03962>.
- Altmeyer, Patrick, Arie van Deursen, et al. 2023. "Explaining Black-Box Models Through Counterfactuals." In *Proceedings of the JuliaCon Conferences*, 1:130. 1.
- Daxberger, Erik, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. 2021. "Laplace Redux-Effortless Bayesian Deep Learning." *Advances in Neural Information Processing Systems* 34.
- Grathwohl, Will, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. 2020. "Your Classifier Is Secretly an Energy Based Model and You Should Treat It Like One." In *International Conference on Learning Representations*.

Hengst, Floris den, Ralf Wolter, Patrick Altmeyer, and Arda Kaygan. 2024.

“Conformal Intent Classification and Clarification for Fast and Accurate Intent Recognition.” <https://arxiv.org/abs/2403.18973>.

Joshi, Shalmali, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. 2019. “Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems.”

<https://arxiv.org/abs/1907.09615>.

Schut, Lisa, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al. 2021. “Generating Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic and Aleatoric Uncertainties.” In *International Conference on Artificial Intelligence and Statistics*, 1756–64. PMLR.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2017. “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR.” *Harv. JL & Tech.* 31: 841. <https://doi.org/10.2139/ssrn.3063289>.

Zagorac, Ivor. 2024. “A Study on Counterfactual Explanations: Investigating the Impact of Inter-Class Distance and Data Imbalance.”