Humans often make poor judgments due to cognitive biases, fatigue, incorrect prior beliefs, or inattentiveness. While AI systems have the potential to support these judgments, humans frequently struggle to collaborate with AI effectively because the same underlying limitations hinder their interaction.

My research develops foundational frameworks for understanding and improving human-AI collaboration. To date, I have demonstrated that common assumptions about AI are flawed – explanations fail as safeguards against algorithmic discrimination, and interpretability increases overreliance on automated systems. Conversely, I have also demonstrated productive pathways forward by developing methods for interpreting reinforcement learning (RL) policies and showing that they can be applied to genuinely improve human decision-making.

Building on these insights, I outline an agenda for two projects: (1) studying user behavior with generative AI systems to understand what interface designs work best, and (2) creating systems that learn from user feedback to personalize AI outputs more efficiently. Together, these projects bridge the gap between AI capabilities and user needs.

To date, my work has fostered international collaborations across machine learning, cognitive science, philosophy, and public policy, with publications at top venues including Machine Learning, Behavior Research Methods, Cognitive Science, and FAccT.

Research Areas: Human-Centered AI, AI for Ethics & Fairness, Interactive ML & Agents

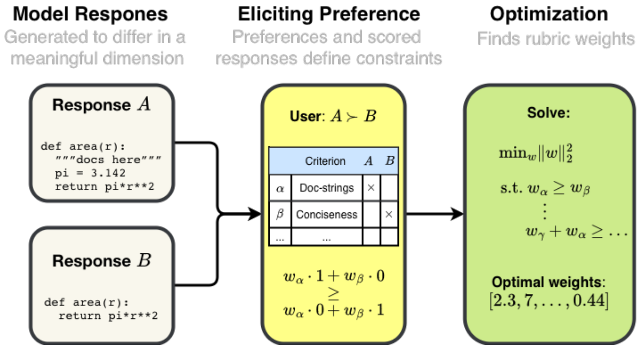
Proposed Research

Area: Human-Centered AI

The gap between AI model performance in controlled evaluations and practical utility in real-world deployment represents a fundamental challenge limiting the impact of current AI systems. Closing this gap requires both (1) better methods for adapting AI to user needs and (2) a deeper understanding of how users interact with these systems. In what follows, I describe two projects that advance these directions.

Specializing LLMs Without Fine-Turning

Language models built for specialization and personalization represent a natural evolution in AI development. Such systems enable efficient code generation suited for specific teams (deployed across companies like Google, Microsoft or IBM) and can democratize access to therapeutic or medical support. However, obtaining specialized models currently relies on computationally expensive retraining or fine-tuning. Recently, researchers proposed to use rubrics – linear scoring functions on features that represent meaningful evaluation criteria (like 'conciseness,' 'doc strings', or 'comments' in code generation) – to score model responses and select optimal outputs. Rubrics have demonstrated success code evaluation and medical education, but their effectiveness depends heavily on rubric weights, which are currently set by experts using heuristics and common sense. As a result, this process is consuming a lot of time and resources.


Model Respones
Generated to differ in a meaningful dimension
Eliciting Preference
Preferences and scored responses define constraints
Optimization
Finds rubric weights
Response A
Response B
User: $A \succ B$
Solve:
$\min_w \|w\|_2^2$
s.t. $w_\alpha \geq w_\beta$
$\vdots$
$w_\gamma + w_\alpha \geq \dots$
$w_\alpha \cdot 1 + w_\beta \cdot 0 \geq w_\alpha \cdot 0 + w_\beta \cdot 1$
Optimal weights: $[2.3, 7, \dots, 0.44]$

```
def area(r):
  """docs here"""
  pi = 3.142
  return pi*r**2
```

```
def area(r):
  return pi*r**2
```

	Criterion	A	B
α	Doc-strings	×	
β	Conciseness		×
...	...		

Fig. 1: Converting user preferences into optimized rubric weights.

In this project, I will automate the process of rubric creation by learning optimal weights from user preferences (see Fig. 1). My approach will associate each rubric criterion $c : R \rightarrow \{0, 1\}$, where $c(r)$ indicates whether response r satisfies criterion c, as a unit vector v_c. I will gather user preferences on pairs of model responses differing in c: $r_c \succeq r'_c$ where $c(r_c) = 1$ and $c(r'_c) = 0$. Each model response r can be then represented as a binary vector $V_r = \sum_{\{c:\ c(r)=1\}} v_c$. This representation allows user preferences to be encoded as linear inequalities: if $r_c \succeq r_c$ then $w \cdot V_{r_c} \geq w \cdot V_{r'_c}$ where w represent the rubric's weights, and use these inequalities to solve an optimization problem. This framework extends naturally to multi-stakeholder alignment. When preference conflicts arise, we could employ selective aggregation and establish partial orderings where preference exists only when it has enough support in the data.

My hypothesis in this project is that automated rubric weight optimization will produce more accurate and consistent evaluations compared to expert-defined heuristic rubrics. Since optimization offers a principled, data-driven approach to rubric creation, this research can significantly speed up rubric generation and offer a scalable solution for creating specialized and personalized models.

Designing for Engagement in Generative AI Models

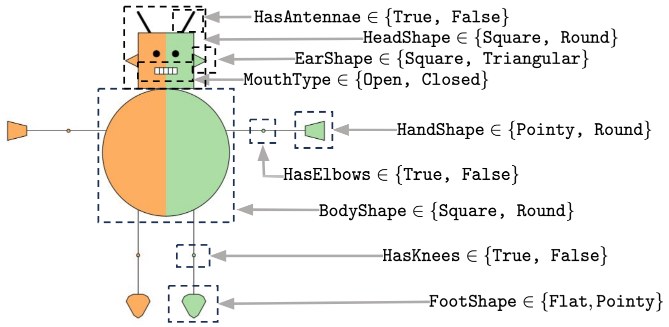

HasAntennae ∈ {True, False}
HeadShape ∈ {Square, Round}
EarShape ∈ {Square, Triangular}
MouthType ∈ {Open, Closed}
HandShape ∈ {Pointy, Round}
HasElbows ∈ {True, False}
BodyShape ∈ {Square, Round}
HasKnees ∈ {True, False}
FootShape ∈ {Flat, Pointy}

Fig. 2: Example robot with features users could specify when generating robots through our proposed generative AI platform. This controlled generation task enables systematic testing of interface design choices with clear success criteria and reproducible experimental conditions.

While generative AI systems have become increasingly powerful, we lack a systematic understanding of how users actually interact with these tools in practice. Current generative AI interfaces offer multiple input modalities, including text descriptions, concept selection, style choices, and reference images. However, we don't know which combinations are most effective, where users struggle, or what causes abandonment. Without empirical evidence of user behavior patterns, interface designers often rely on intuition rather than data, which limits their ability to optimize the user experience.

In this project, drawing on my experience in creating experimental platforms for studying interpretability [SGU24] and discrimination detection [SDU25], I will develop a platform for testing how users engage with various generative AI interface designs. To design the platform, I will extend my existing robot classification environment [SGU24], by enabling users to request images of robots with specific characteristics. This controlled environment provides clear success metrics (achieving specific robot features) and eliminates confounding variables that complicate open-ended creative tasks (user intent, domain expertise, etc.). Through a series of studies, I will investigate which design choices optimize task completion speed, output quality, and sustained engagement. Initial studies will focus on two key dimensions: (1) concept specification methods (natural language vs. structured controls vs. example-based) and their effects on task completion and output quality, and (2) iteration workflows (real-time editing vs. generate-then-refine vs. conversational steering) and their effects on engagement and abandonment rates. The platform will capture detailed interaction logs, including the time spent, revision sequences, and points of abandonment. My hypotheses are that: (1) structured controls will enable faster convergence for specific targets while natural language will better support exploratory tasks, and (2) real-time editing will reduce abandonment by providing immediate feedback. This research aims to establish evidence-based design principles for generative AI interfaces, transitioning the field to data-driven design decisions.

Past Research — Areas: AI for Ethics & Fairness, Interactive ML & Agents

Effect of Interpretability on Human-AI Collaboration [SGU24]

A core assumption in human-AI interaction is that interpretable models lead to better human decision-making. The rationale is that knowing how the model maps inputs to outputs helps identify model errors, biases, or spurious correlations, and realize when the user has side information. It is a problematic assumption to validate because the results may not generalize across use cases (e.g., they can differ in clinical vs. judicial decision-making), are influenced by people's prior beliefs, and depend on the comparison to optimal decisions under available information.

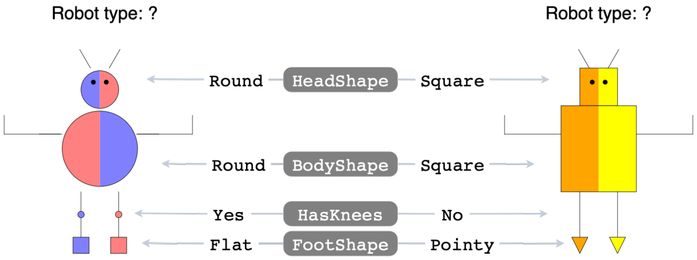


Fig. 3: In our platform, a model predicts robot type based on four binary characteristics, ($\mathbb{1}[\texttt{HeadShape} = \texttt{Round}]$, $\mathbb{1}[\texttt{Bodyhape} = \texttt{Round}]$, $\mathbb{1}[\texttt{Knees} = \texttt{Yes}]$, $\mathbb{1}[\texttt{FootShape} = \texttt{Pointy}]$). We investigate whether, once the model is made interpretable and explicitly shows how robot features predict robot type, this helps people make more accurate predictions themselves.

Linear Classifier $\hat{y} = 1 \ \text{ if } \ 1.74x_1 + 1.10x_2 \geq 2.84$

$\Updownarrow$

Scoring System $\hat{y} = 1 \ \text{ if } \ x_1 + x_2 \geq 2$

$\Updownarrow$

Boolean Rule $\hat{y} = x_1 \ \text{ AND } \ x_2$

Fig. 4: In our platform, we represent models in interpretable formats like a linear function (top) or a scoring system (a linear function with integer coefficients that can be treated as a weighted checklist; middle). We found that the expert's reliance on the models varies with the format, resulting in different degrees of overreliance.

In this project, I co-ideated a fully controllable setup (robot type prediction, as shown in Fig. 3), computed all linearly separable models in the environment, and designed quantitative metrics for model comparison to measure the effect of interpretable models on decision-making (models depicted in Fig. 4). In my setup, I can set the model's accuracy, format, and its relationship to the model that describes the user's beliefs, among other parameters, and define the default reliance on the model's predictions. I used this setup to demonstrate that humans overrely on interpretable models, irrespective of their accuracy, and the level of reliance changes with the chosen model format (as shown in Fig. 4) [SGU24]. Currently, we are conducting additional studies and plan to submit our work to an HCI venue (CHI) and a general science journal (PNAS).

Improving Human Decision-Making Through Automatically Discovered Decision Aids [SBL21; BSOL22]

A fundamental challenge in human-AI interaction is that users often cannot harness the benefits from AI systems because they cannot understand them. Reinforcement learning (RL) is one area where insights from AI could help both experts (e.g., in sepsis treatment) as well as the general public (to learn clever heuristics for decision-making). The primary issue is that RL generates opaque, stochastic black-box policies that offer no insight into the underlying decision-making logic.

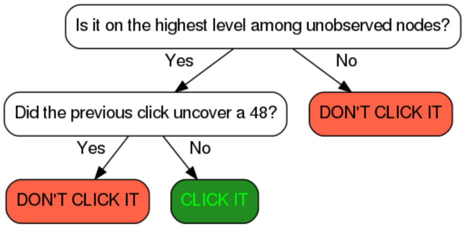

Is it on the highest level among unobserved nodes?
Yes
No
Did the previous click uncover a 48?
DON'T CLICK IT
Yes
No
DON'T CLICK IT
CLICK IT

Fig. 5: Automatically discovered heuristic in a tree-graph environment where nodes encode outcomes of actions (the higher, the more long-term). The strategy assumes inspecting all long-term outcomes until the best one is found.

To elucidate the end-product of RL, I developed a Bayesian imitation learning algorithm that outputs interpretable descriptions of policies as flowcharts (as in Fig. 5) [SBL21]. My approach combines advances in imitation learning and program induction with a novel clustering method for identifying subsets of demonstrations that can be accurately described by simple, high-performing decision rules. The key innovation is that my algorithm can handle cases where traditional methods fail, i.e., when the created domain-specific language is insufficient to describe the whole set of demonstrations or when policies exhibit idiosyncratic behaviors that cannot be captured by available language. Through extensive behavioral experiments, I demonstrated that humans can successfully understand and apply the flowcharts generated by my method, significantly improving their strategies across different sequential decision problems. Most importantly, my approach proved more effective than conventional training methods that rely on performance feedback, establishing that AI-discovered strategies can be successfully transferred to human decision-makers through proper interpretability tools.

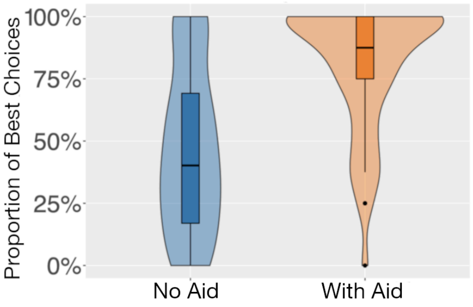


Fig. 6: Proportion of choices for the mortgage with the lowest interest rate when participants can use the decision aid versus when they cannot.

I then extended my work to create practical interventions that help humans apply AI-discovered heuristics in naturalistic settings [BSOL22]. I focused on promoting far-sightedness – countering human tendency to prefer immediate gains over long-term rewards – in two tasks: planning a road trip and choosing a mortgage. My key finding was that the format of the intervention matters critically, as procedural instructions (step-by-step guides) led to significantly better outcomes than static flowcharts. This finding motivated my primary technical contribution — an algorithm that transforms flowcharts into procedural instructions using linear temporal logic. My empirical contribution was that decision aids generated with this algorithm succeed in promoting better decisions (see **??**). This work demonstrates that while humans can benefit from AI-discovered strategies, success depends on translating these strategies into formats that align with how people naturally think and process information.

Evaluating Explanations for Algorithmic Discrimination [SDU25]

One of the challenges in human-AI collaboration is that technological progress outpaces empirical validation. This is especially important for research in explainable AI as regulatory frameworks increasingly mandate model explanations as safeguards against algorithmic discrimination (see the example in Fig. 7), assuming humans can reliably use these explanations to detect unfair predictions. However, this assumption had never been rigorously tested under controlled conditions, be-

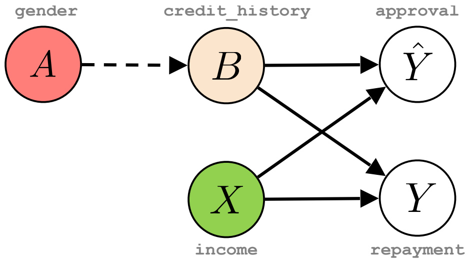


Fig. 7: Causal diagram for discrimination detection. For example, in loan approval predictions ($\hat{Y}$), the model uses an individual's income (X) and credit history (B) as inputs. Gender (A) could affect credit history due to differences in credit scores or the intensity of credit usage found between men and women.

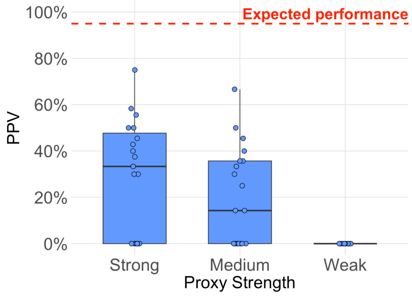


Fig. 8: Reliability of discrimination detection with explanations. We expect people to spot $\geq 95\%$ of discrimination (red line). Instead, people retrieve only 50% of all unfair predictions, irrespective of how strong the proxy for the protected attribute is.

cause discrimination detection is inherently a probabilistic task without a clear ground truth. Detection is also prone to human errors tied to flawed prior knowledge.

In this work, I developed synthetic controlled conditions to test these assumptions, ultimately conveying the limitations of explanations for assessing fairness [SDU25]. I created a framework for judging the fairness of predictions at an instance level through counterfactual fairness theory, where a prediction is (δ-)counterfactually fair if changing the protected attribute can change the probability of obtaining the same prediction (by at most δ). I created a synthetic environment where I could control failure modes of discrimination detection unrelated to explanations: whether people know the proxies, their strength, the protected class for predictions they study, whether they know how to use explanations to support claims, and whether explanations hide relevant information.

I utilized my framework to empirically demonstrate that explanations are unreliable for assessing discrimination. Even with perfect knowledge of proxy strength, protected class participants retrieve between 40% and 70% of the discriminatory predictions (see Fig. 8), and systematically label 30% of all fair predictions as discriminatory, revealing their incorrect assumptions about proxy identity. This work provides evidence that current regulatory approaches, which rely solely on explanations, are insufficient, underscoring the broader need to validate assumptions about human-AI collaboration empirically. To date, it has drawn interest from governmental agencies in the US, the Netherlands, and Switzerland.

Evaluating Concept Bottleneck Models [CKSSU25]

Concept bottleneck models (CBMs) are deep learning models that detect human-interpretable concepts in the input data and use them to make final predictions (see Fig. 9). This bottleneck allows humans to inspect the detected concepts and correct them to improve model performance. The greatest barrier to adopting these systems in practice is costly concept annotation. This investment would be justifiable if practitioners could thoroughly evaluate whether CBMs work for their specific use case. Yet due to the lack of reliable benchmarks that offer controlled variation, practitioners cannot make informed deployment decisions. In this project, we introduced synthetic benchmarks for two critical applications for CBMS: automating human work (where we used validation of Sudoku boards) and supporting human decision-making (where we used robot classification task from [SGU24]).

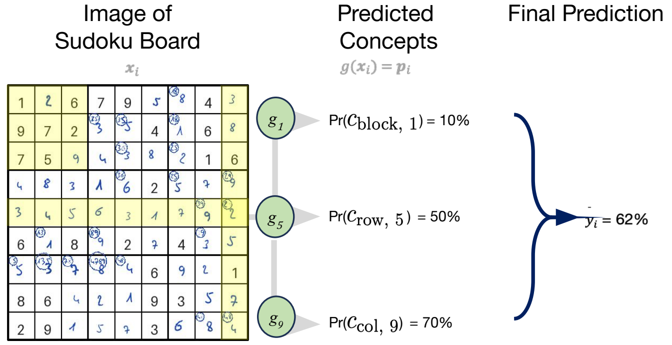


Fig. 9: Concept bottleneck models detect high-level features from the data, and then aggregate them to obtain the final prediction.

Our experiments revealed that while CBMs can outperform standard neural networks, they are highly sensitive to realistic deployment challenges. Results showed that overly granular concept sets reduced accuracy below the standard network performance (from 97.7% to 87.1% accuracy), concept annotation noise eliminated intervention benefits entirely (reducing gains from +16.3% to 0% accuracy), and expert constraints intended to improve model alignment actually degraded intervention effectiveness (reducing accuracy gains by over 10%). This work provides practitioners with the first systematic framework for evaluating when investments in CBMs are justified, illustrating how we can improve human-AI collaboration by improving evaluation of existing methods.

References

[BSOL22] Frederic Becker et al. "Boosting human decision-making with AI-generated decision aids". In: **Computational Brain & Behavior** 5.4 (2022), pp. 467–490. **url**: https://link.springer.com/article/10.1007/s42113-022-00149-y.

[CKSSU25] Harry Cheon et al. "Measuring What Matters: Synthetic Benchmarks for Concept Bottleneck Models". In: **In Submission** (2025). **url**: https://www.jskirzynski.com/static/concept_benchmark.pdf.

[SBL21] Julian Skirzyński, Frederic Becker, and Falk Lieder. "Automatic discovery of interpretable planning strategies". In: **Machine Learning** 110.9 (2021), pp. 2641–2683. **url**: https://link.springer.com/article/10.1007/s10994-021-05963-2.

[SDU25] Julian Skirzyński, David Danks, and Berk Ustun. "Discrimination Exposed? On the Reliability of Explanations for Discrimination Detection". In: **Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency**. 2025, pp. 2554–2569. **url**: https://dl.acm.org/doi/pdf/10.1145/3715275.3732167.

[SGU24] Julian Skirzyński, Elena Glassman, and Berk Ustun. "On Interpretability and Overreliance". In: **Interpretable AI: Past, Present and Future NeurIPS Workshop**. 2024. **url**: https://openreview.net/pdf?id=0w7JBZibDc.