People 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 shown productive pathways forward by developing methods for interpreting reinforcement learning (RL) policies and demonstrating that it could be applied to genuinely improve human decision-making.

Building on these insights, I outline an agenda for two projects: (1) extracting interpretable concepts from ML models to enable targeted user feedback, and (2) studying user engagement with generative AI interfaces to establish evidence-based design principles. Together, these projects advance user control over AI systems by both building better control mechanisms and understanding how users want to interact with AI.

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

AI systems are expanding rapidly from expert domains to general consumer use. This shift requires developing new insights, as general users have different needs, capabilities, and contexts from trained specialists. I address this challenge through two complementary goals: (1) developing methods that enable meaningful user control over AI behavior; (2) understanding what controls users actually need through systematic study of their interactions. In what follows, I describe two projects that advance these directions.

User Control Through Concepts Extracted from Traditional ML Models

As AI enters mainstream use, people find themselves unable to provide targeted feedback that would change the model's predictions in meaningful and expected ways. Current approaches offer only crude binary ratings, such as likes/dislikes, that miss nuanced preferences or wholesale model retraining that lacks precision and clarity about what actually changed. Concept architectures partially address this limitation by allowing users to interact with interpretable concepts. They represent the model's intermediate outputs as high-level concepts and train the model to use these concepts to predict the final output. This allows users to intervene on the concepts at inference time, making outputs "more formal", "less technical", or adjusting other high-level attributes. The main problem is that these methods are limited to differentiable algorithms like neural networks, leaving traditional ML models like Learning-to-Rank systems without interpretable user control.

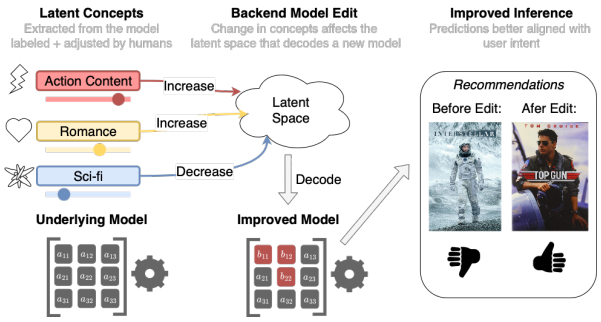

Fig. 1: Parameterization of machine learning models encodes latent, interpretable concepts that shape how predictions are made. For instance, a movie recommender systems may use concepts for the amount of action, romance and sci-fi motives. Users can directly adjust a model by interacting with these concepts, making it update its internal structure. This would produce outputs that better reflect user intent.

To remedy that lack of control over non-differentiable models, I will first design methods to identify concepts they use for prediction. One promising approach builds on recent advances in concept generation for deep learning. This approach extracts concepts as vectors that span the encoding matrix of a neural network layer. I aim to adapt this approach to traditional ML models by constructing feature relationship matrices weighted by model parameters (e.g., feature correlations weighted by model coefficients). I will apply PCA to such matrices to extract orthogonal directions, which I will treat as latent concepts. I will then re-express the model's parameters in this concept space, yielding a concept-based model that makes identical predictions but operates on interpretable conceptual dimensions rather than raw features.

I will build on existing approaches where humans learn to understand concepts by observing their effects on model predictions. Specifically, people will examine predictions from models using only individual concepts (with all other concept weights set to zero) across representative examples, identifying patterns in how each isolated concept responds to different inputs. A key limitation of such observational learning is that people anchor to initial impressions or interpret concepts through flawed prior beliefs. To address this, I will adapt debiasing techniques from my prior work on teaching

interventions [SBL21; BSOL22], specifically developing a concept-specific version of the competing explanations intervention. Rather than accepting their first interpretation, people will generate and evaluate multiple candidate explanations for what each concept represents before selecting the most supported interpretation. This promotes deeper, structural reasoning about concept meaning.

My hypotheses in this project are that:

1. Principal components extracted from model-informed feature matrices can be meaningfully interpreted by humans through observing predictions from models using only individual components
2. Generating multiple hypotheses about concept meanings counters cognitive biases and increases labelers' agreement

This research enables more effective human-AI collaboration across multiple domains (e.g., see Fig. 1).

1. In consumer applications, my concept extraction methods would identify the key dimensions driving recommendations (e.g., "recent history" vs "declared preferences"), present these concepts to users, and allow users to adjust concept weights to better align with their true preferences.
2. In expert settings, concepts may be used to discover "superhuman" knowledge not represented in human prediction data. e.g., by retrieving such concepts and validating their meaning with human experts (like medical professionals).
3. Concepts may also be used by developers who want to steer the model towards particular prediction patterns by performing targeted interventions, e.g., emphasizing safety-related concepts in autonomous systems.

Designing for Engagement in Generative AI Models

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—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 about user behavior patterns, interface designers rely on intuition rather than data, limiting their ability to optimize 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. 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 time spent, revision sequences, and abandonment points. 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 will establish evidence-based design principles for generative AI interfaces, moving the field from intuition-driven to data-driven design decisions.

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

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 main issue is that RL generates opaque, stochastic black-box policies that provide no insight into the underlying decision logic.

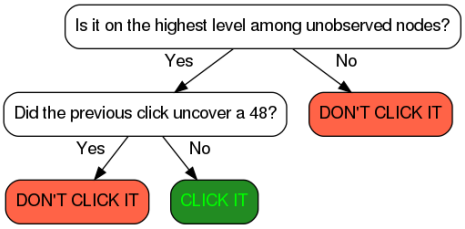

Fig. 2: Automatically discovered heuristic in a tree-graph environment where nodes encode numeric outcomes of actions (the higher, the more long-term). The strategy assumes inspecting all long-term outcomes until the best one is found.

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. 2) [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 large behavioral experiments, I demonstrated that people 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.

I then extended my work to create practical interventions that help people apply AI-discovered heuristics in naturalistic settings [BSOL22]. I focused on promoting far-sightedness – countering people's 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 motivated my main technical contribution – an algorithm that transforms flowcharts into procedural instructions via linear temporal logic. My empirical contribution was that decision aids generated with this algorithm succeed in promoting better decisions (see Fig. 4). This work demonstrates that while humans can benefit from AI-discovered strategies, but success depends on translating these strategies into formats that align with how people naturally think.

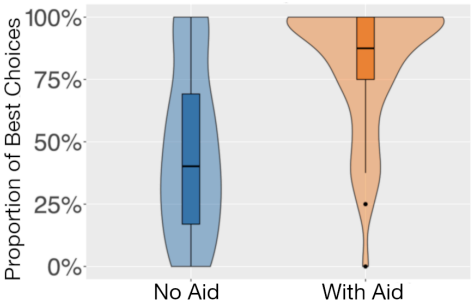

Fig. 3: Proportion of choices for the mortgage with the lowest interest rate when participants can use the decision aid versus when they cannot.

Automatically Discovering Human Planning Strategies [SJL24]

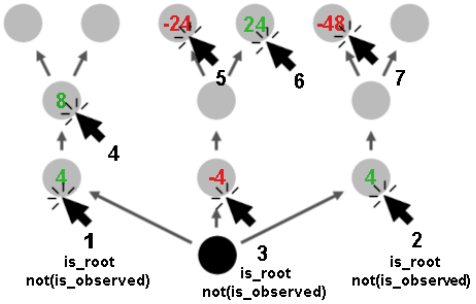

Fig. 4: Sequence of externalized planning operations (clicks) where the task is to find the most rewarding route from the black node to the top, uncovering the fewest possible nodes. Numbers denote the ordering of operations (clicks), whereas the predicates denote relevant elements of the Domain Specific Language active when using a given operation.

Traditional approaches to understanding how people make decisions and plan require months of manual analysis, creating a significant bottleneck in research. In this project, I developed an automated method for discovering and describing human planning strategies from behavioral experiments where participants' decision-making processes are externalized through their interactions with a computer interface. My approach uses an expectation-maximization algorithm to cluster sequences of human behaviors based on their statistical similarity, and then applies my algorithm for interpreting reinforcement learning policies to describe each cluster. The clustering groups similar behavioral sequences (e.g., sequences of clicks as in Fig. 4), while the interpretation step analyzes which elements of a Domain Specific Language are active in each sequence to generate procedural descriptions that can imitate the observed behaviors with high fidelity.

The benchmark experiments demonstrated that this framework reduced the analysis time from 120 days to 20 days, while rediscovering over 50% of the most frequently used strategies. This 6-fold acceleration shows how AI can augment human scientific capabilities, rather than replace them, enabling researchers to process larger datasets more efficiently and direct their efforts elsewhere.

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. 5), assuming humans can reliably use these explanations to detect unfair predictions. However, this assumption had never been rigorously tested under controlled conditions, because discrimination detection is inherently a probabilistic task without clear ground-truth. Detection is also prone to human errors tied to flawed prior knowledge.

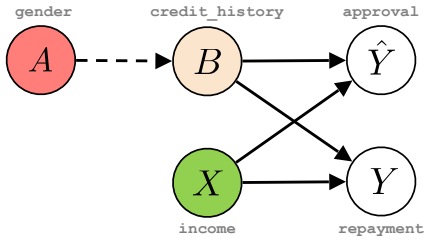

Fig. 5: Causal diagram for discrimination detection. For example, in loan approval predictions ($\hat{Y}$), the model uses an individual's income (X) and credit history (B) as inputs. Gender (A) could affect credit history due to differences in credit scores or the intensity of credit usage found between men and women.

In this work, I developed synthetic controlled conditions where I could test these assumptions, to ultimately convey the limitations of explanations for assessing fairness [SDU25]. I developed 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 used my framework to empirically show that explanations are unreliable for judging discrimination. Even under perfect knowledge of proxy strength and protected class participants retrieve between 40% to 60% of the discriminatory predictions (see Fig. 6), and systematically label 30% of all fair predictions as discriminatory, showing their incorrect assumptions about proxy identity. This work provides evidence that current regulatory approaches relying on explanations alone are insufficient, highlighting the broader need to empirically validate assumptions about human-AI collaboration. To date, it has drawn interest from governmental agencies in the US, the Netherlands and Switzerland.

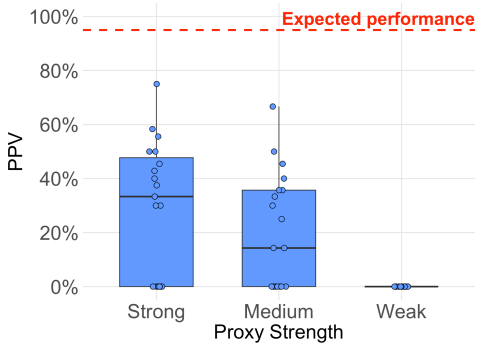

Fig. 6: Reliability of discrimination detection with explanations. We expect people to spot $\geq$ 95% of discrimination (red line). Instead, people retrieve only 50% of all the unfair predictions irrespective of how strong the proxy for the protected attribute is.

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 spotting model errors, biases, spurious correlations, or realize when the user has side-information. It is a difficult assumption to validate because the results might not generalize across use cases (e.g., they can be different in clinical vs. judicial decision-making), they are affected by people's prior beliefs, and they depend on the comparison to optimal decisions under available information.

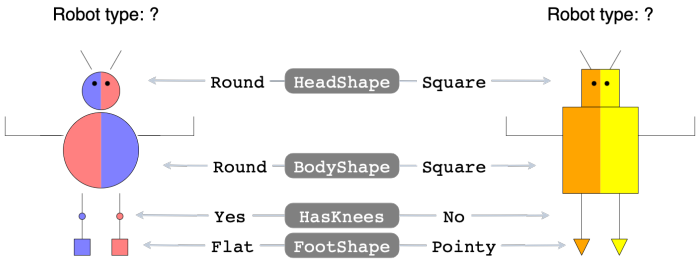

Fig. 7: In our platform, a model predicts robot type based on 4 binary characteristics, ($\mathbb{1}$[HeadShape = Round], $\mathbb{1}$[Bodyhape = Round], $\mathbb{1}$[Knees = Yes], $\mathbb{1}$[FootShape = Pointy]). We study if once the model is made interpretable and explicitly shows how robot features predict robot type this helps people make better predictions themselves.

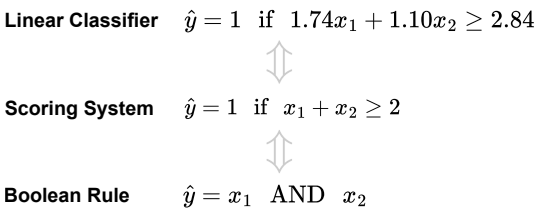

Fig. 8: In our platform, we represent models in interpretable formats like a linear function (top) or a scoring system (a linear function with integer coefficients that can be treated as a weighted checklist; middle). We found that expert's reliance on the the models changes with the format, leading to different degrees of overreliance.

In this project I co-ideated a fully controllable setup (robot type prediction from Fig. 7), 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 as in Fig. 8). In my setup, I can set the model's accuracy, its format, its relationship to the model describing the user's beliefs, etc., and define benchmark reliance on the model's predictions. I used this setup to show that people overrely on interpretable models irrespective of their accuracy, and the level of their reliance changes with the chosen model format (as in Fig. 8) [SGU24]. Currently, we are running more studies and plan to submit the work to an HCI venue (CHI) and a general science journal (PNAS).

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.

[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.

[SJL24] Julian Skirzyński, Yash Raj Jain, and Falk Lieder. "Automatic discovery and description of human planning strategies". In: **Behavior Research Methods** 56.3 (2024), pp. 1065–1103. **url**: https://link.springer.com/article/10.3758/s13428-023-02062-z.