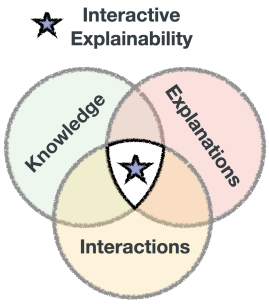
Artificial Intelligence (AI) systems that use large language models (LLMs), like GPT-3, have made incredible progress in performing complex tasks that humans otherwise do effortlessly. Current AI can reason about the world, follow instructions during navigation, and generate rich multimedia content with remarkable efficiency. However, these models often ignore the long tail of the information, such as user preferences, cultural nuances, and domain knowledge, blocking end users from reaping the full benefit of their scale. My research stands at the heart of the following question:
what can we gain by reinventing the AI systems to begin from users?



Users continuously seek to access information to make a decision (e.g., purchase), reduce uncertainty (e.g., understanding health symptoms), or accomplish a task (e.g., navigation). First, **current systems** which produce machine predictions to assist a user (e.g., recommendations) **are often opaque and not interpretable** for practical use. Second, **current explainable systems primarily render low-level interpretations** (e.g., just showing the most toxic words to explain a tweet as toxic), failing to address the complex reasoning associated with social and historical context. Third, **existing AI systems only solve a narrow class of problems** (that have a unique *correct* answer) and are not applicable where outputs are personalized, or decisions are subjective (e.g., computational ethics).

An assistive AI system must be aware of the surrounding world (**relevant**), produce consistent and faithful explanations (**trustworthy**), and align with the user's preferences (**adaptive**) [1]. My research concisely addresses the above through **interactive explainability**, realized via three interwoven pillars[1]:

1. **Knowledge**: Discovering and deducing knowledge from context using external resources or via clarifying conversations [2, 3, 4, 5, 6, 7, 8],
2. **Explanations**: Enhancing machine predictions and their explanations by aligning the model's reasoning process with world and personal knowledge [9, 10, 11, 12, 13],
3. **Interactions**: Enabling users to critique and update model beliefs to align predictions and explanations to their personal, social, and subjective contexts [14, 15].

1 Knowledge Grounding for Relevant AI

Knowledge in its surface form is either propositional (facts) or the perception of skills, objects, and events (commonsense). However, required contextual knowledge greatly varies based on the **personal preferences of the user**, subjectivity associated with the context, and availability of resources to acquire such knowledge.

When AI models act as *experts* in a knowledge-seeking scenario (e.g., seeking recommendations, obtaining explanations), **they often ignore subjective preferences or become limited to the knowledge they are trained on**. In [2, 3], we showed that a model trained on persona-grounded conversations could not deduce implicit knowledge from the dialog context. For example, humans can easily infer that *if someone likes hiking, they may love nature or want to be fit*; in contrast, traditional generative models fail to acknowledge it.

To remedy this, we explored a **training-time knowledge augmentation** paradigm that expands a textual context into possible inferential knowledge and then augments therein [2]. Critically, we use a retrieve-generate framework that first uses external commonsense knowledge graphs (e.g., ATOMIC [16]), web-

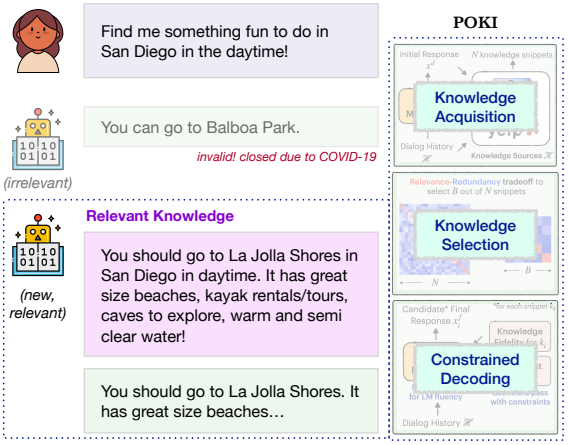


Figure 1: **Po**st-hoc **K**nowledge **I**njection (**POKI**) [4] in a dialog model that was trained on limited knowledge

[1] A complete list of publications and media coverage articles are at my website.

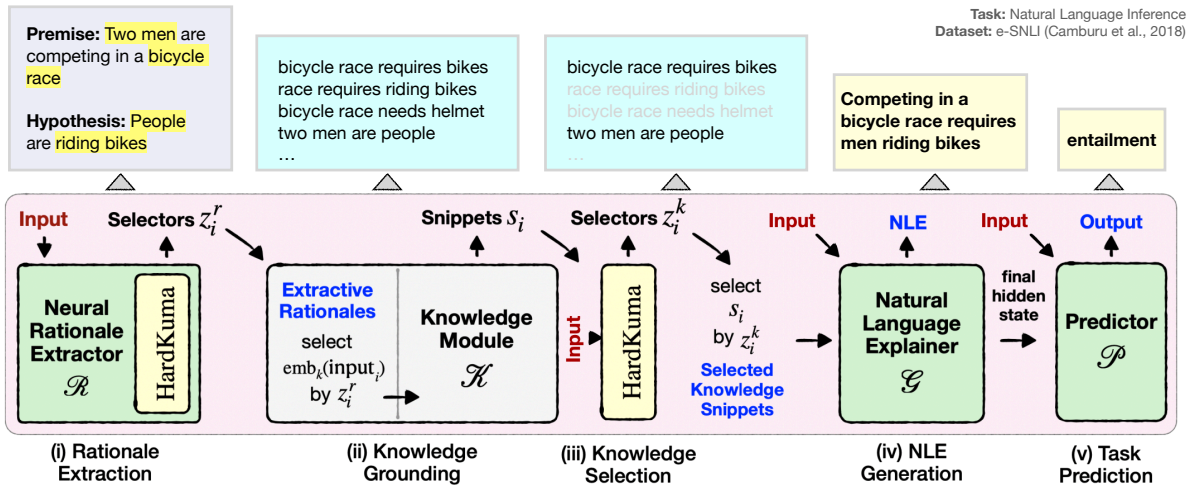


Figure 2: **REXC** [9] bridges extractive rationales and abstractive NLEs using background knowledge to produce quality explanations and accurate predictions.

scale corpora (e.g., Yelp Reviews), or generative models (e.g., GPT3, COMET [17]) as knowledge sources and then augments the retrieved knowledge in the dialog model using variational learning. We found that humans predominantly prefer our generated responses as they are highly diverse, attributable, and controllable with input persona.

Training-time knowledge augmentation requires continuous fine-tuning to keep models up-to-date, resulting in higher carbon footprints. We developed post-hoc approaches [3, 4] to knowledge acquisition and injection for existing dialog models to make the process more lightweight and greener. It also applies when the underlying model's parameters are not accessible or updatable. This time, we retrieved the additional knowledge and used a *post-hoc* gradient-based method (POKI) to inject new knowledge into a generated dialog response both **at inference time**. We showed (see Figure 1) that when we injected up-to-date knowledge (e.g., post-COVID travel regulations) in an existing dialog model (let's say, trained in pre-COVID time), users could efficiently (re)use it to reach their conversational goals (e.g., planning their travel in 2022).

We further realized that knowledge-augmentation techniques detailed so far will still be ineffective when the context is ambiguous. For example, suppose a user is looking for a *travel recommendation*. Several nuances (e.g., number of travelers, location, and transport preferences) can change the knowledge requirement given the same context. We later developed a question generation framework to estimate the missing information, pose relevant and **useful questions to reduce ambiguity** [5], and gather knowledge that aligns with user preferences.

Finally, we tested our knowledge-augmentation techniques **at a scale of millions of users**. We showed a **65% improvement in user-satisfaction** and more than **180% increase in user-engagement** by making a dialog agent knowledge-aware and up-to-date [18][2].

2 Generating Explanations 🔍 for Trustworthy AI

Machines often perform better than an average human being in many tasks by solving them *differently* than humans. Hence, it is crucial to understand the model's underlying representation for better scientific understanding [10] and improved trust [9]. To this end, user-centric models must produce comprehensive, personalized explanations 🔍 attributable to world knowledge.

I primarily worked with expressive forms of explanations, such as rationales (predictive parts of input) or natural language explanations (NLEs), that could provide more accessibility to users and cover subjective contexts. Upon investigation, we found that existing explainable models often lack background knowledge, affecting task performance and explanation quality. In [9], we showed **adequate knowledge grounding** for three natural language tasks and two vision-language tasks could improve the quality of the explanations to be state-of-the-art (in REXC, see Figure 2). Additionally, we achieved the best task

[2]Presented in the finals of Amazon Alexa Prize, as 1 of 10 teams from 300 international participants, and awarded $250,000.

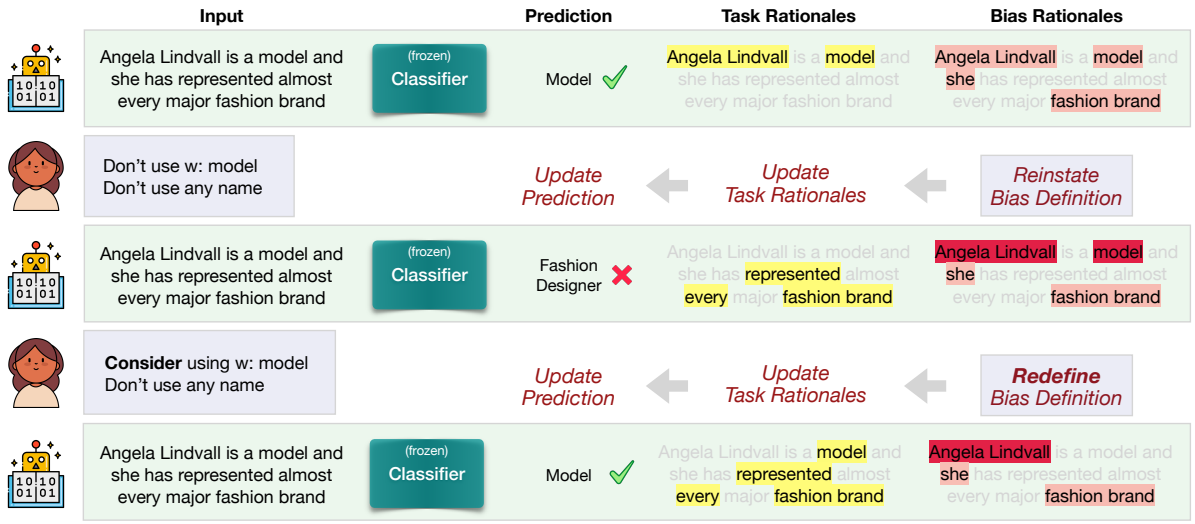


Figure 3: **INTERFAIR** [15], a new interactive paradigm of controllable debiasing. Users can update the model's belief about sensitive information to mitigate bias while maintaining the task performance.

performance across all equivalent explainable models—indicating that REXC **closes the critical gap between task performance and explainability**.

Knowledge-grounding to improve explanation quality further gave rise to several emergent properties of the explanations: **factuality**, **robustness**, and **faithfulness—critical for enabling the user to take actions based on the explanations**. In [9], we observed that generated NLEs exhibit a high degree of faithfulness; a similar observation was made for T5-based self-rationalizing models [19]. Similarly, these models are more robust to knowledge-based adversarial attacks (e.g., change of entity, negations) than not knowledge-grounded models [20]. This effect is more pronounced in domain-specific applications (e.g., e-commerce), where several state-of-the-art NLEs generation models succumb to the issue of hallucination [11]. We partially addressed this issue by evidence-grounding the NLEs and generating significantly more factual NLEs compared to previous state-of-the-art [13].

We merely understand black-box models and how they encode complex social contexts into model parameters. We find that **rationales are useful in exposing the model's understanding of complex social contexts**. We find that traditional models that address the issue of incorrect social understanding (e.g., exposure bias) are often opaque and do not provide reliable explanations for why they are unbiased. Instead, these models perform debiasing too harshly and disregard task performance. In [12], we show how to control the bias exposure by constraining the task rationales to be minimally biased to retain the original task performance.

3 Engaging in Interactions for Adaptive AI

Despite our best efforts to make AI systems knowledgeable and explainable, data around us will still be inherently biased and limited by its origin. The models we build will be less than perfect. Moreover, humans also constantly modify their expectations from the AI models. Hence, **the gold solution remains to make the user a part of the learning paradigm**. To this end, designing, building, and evaluating interactive models can propel our progress toward anthropomorphic AI systems[3].

Given an input, most ML models produce a static, one-time prediction (and explanations). **Users struggle to meaningfully express individual preferences due to the static nature of ML models**. We considered a recommender system where the user needs to continuously refine the model's prediction (here, the recommended items) according to their preference. We found that explanations provide a useful gateway for users to give feedback. Subsequently, the model can update its prediction by changing its explanations according to the feedback. We modeled this interaction as a *critiquing*[4] process and showed that critiquing improves the model performance and user satisfaction [14]. We also persisted

[3]My proposal was recognised by Adobe Research Fellowship (2022) and Qualcomm Innovation Fellowship (2020).
[4]Our work on *critiquing* has been recognized as 2022 Highlights of ACM RecSys.

the critiques in model parameters so that the model can proactively use them on unseen data points.

We extended our critiquing framework to improve debiasing (introduced in Section 2) performance bringing users into the loop (INTERFAIR). We observed that the trade-off between task performance and bias mitigation greatly varies between users [21] and is often hard to achieve via purely learning from data [22]. Figure 3 shows how users can modify the amount of bias (e.g., towards gender) in model explanations to balance bias mitigation and prediction accuracy.

Updating the model's belief during test time for improved predictions can be extended to any model capable of producing explanations for its predictions. This model debugging process can also be viewed as **teaching AI models**. Our ongoing work on machine teaching [23] to improve the model's (e.g., GPT-3) capability of better domain understanding (e.g., science) has promise for building never-ending learning systems to improve continually over time.

Conclusion: In summary, my research takes a user-centric approach to achieve subjectivity and personalization in AI models. Our explainable interactive systems will positively impact end-user behavior by accelerating the growth of real-world systems. Critiquable systems can address learning differences among users, build skills, and reduce ambiguity in communication, among many other societal impacts. Attributable explanations can aid users with more agency to make high-stake decisions addressing the issue of AI over-trust. While current AI systems struggle to be accessible to non-expert users, the next-generation AI systems powered with knowledge  , explanations  , and interactions  will successfully bridge the gap between humans and AI.

References

[1] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 2019.

[2] **Bodhisattwa Prasad Majumder**, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *EMNLP*, 2020.

[3] **Bodhisattwa Prasad Majumder**, Taylor Berg-Kirkpatrick, Julian McAuley, and Harsh Jhamtani. Unsupervised enrichment of persona-grounded dialog with background stories. In *ACL*, 2021.

[4] **Bodhisattwa Prasad Majumder**, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. Achieving conversational goals with unsupervised post-hoc knowledge injection. In *ACL*, 2022.

[5] **Bodhisattwa Prasad Majumder**, Sudha Rao, Michel Galley, and Julian McAuley. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. In *NAACL-HLT*, 2021.

[6] **Bodhisattwa Prasad Majumder**, Shuyang Li, Jianmo Ni, and Julian McAuley. Generating personalized recipes from historical user preferences. In *EMNLP*, 2019.

[7] Huanru Henry Mao, **Bodhisattwa Prasad Majumder**, Julian McAuley, and Garrison Cottrell. Improving neural story generation by targeted common sense grounding. In *EMNLP*, 2019.

[8] **Bodhisattwa Prasad Majumder**, Shuyang Li, J. Ni, and Julian McAuley. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *EMNLP*, 2020.

[9] **Bodhisattwa Prasad Majumder**, Oana Camburu, Thomas Lukasiewicz, and Julian McAuley. Knowledge-grounded self-rationalization via extractive and natural language explanations. In *ICML*, 2022.

[10] **Bodhisattwa Prasad Majumder**, Navneet Potti, Sandeep Tata, J. Wendt, Qi Zhao, and Marc Najork. Representation learning for information extraction from form-like documents. In *ACL*, 2020.

[11] Zhouhang Xie, Julian McAuley, and **Bodhisattwa Prasad Majumder**. On faithfulness and coherence of language explanations for recommendation systems. *CoRR*, 2022.

[12] Zexue He, Yu Wang, Julian McAuley, and **Bodhisattwa Prasad Majumder**. Controlling bias exposure for fair interpretable predictions. *Findings of EMNLP*, 2022.

[13] Zhouhang Xie, Sameer Singh, Julian McAuley, and **Bodhisattwa Prasad Majumder**. Factual and informative review generation for explainable recommendation. *CoRR*, 2022.

[14] Shuyang Li, **Bodhisattwa Prasad Majumder**, and Julian McAuley. Self-supervised bot play for transcript-free conversational recommendation with rationales. In *ACM RecSys*, 2022.

[15] **Bodhisattwa Prasad Majumder**, Zexue He, and Julian McAuley. InterFair: Debiasing with natural language feedback for fair interpretable predictions. Under review.

[16] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *AAAI*, 2019.

[17] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: commonsense transformers for automatic knowledge graph construction. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *ACL*, 2019.

[18] **Bodhisattwa Prasad Majumder**, Shuyang Li, Jianmo Ni, Henry Mao, Sophia Sun, and Julian McAuley. Bernard: A stateful neural open-domain socialbot. *Alexa prize proceedings*, 2020.

[19] Sarah Wiegreffe, Ana Marasovic, and Noah A. Smith. Measuring association between labels and free-text rationales. In *EMNLP*, 2021.

[20] Myeongjun Jang, **Bodhisattwa Prasad Majumder**, Julian McAuley, Thomas Lukasiewicz, and Oana-Maria Camburu. Know how to make up your mind! Adversarially detecting and remedying inconsistencies in natural language explanations. Under review.

[21] Mohammad Yaghini, Andreas Krause, and Hoda Heidari. A human-in-the-loop framework to construct context-aware mathematical notions of outcome fairness. In *AIES*, 2021.

[22] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AIES*, 2018.

[23] Bhavana Dalvi, Oyvind Tafjord, and Peter Clark. Towards teachable reasoning systems. *CoRR*, 2022.