Reconciling AI Representations and User Mental Models

Angie Boggust

Artificial intelligence (AI) models promise to amplify human capability by serving as extensions of people's cognition and creativity. Yet, models often learn representations of the world that conflict with their users' mental models, leading to unexpected and harmful interactions. Clinical models violate their physicians' medical principles, AI assistants neglect their blind users' needs, and creative tools contradict their artists' cultural narratives. These epistemic conflicts stem from today's human-AI alignment paradigm, where models are optimized for general human norms and deployed as opaque tools. As a result, models fail to reflect users' domain-specific expectations, and users lack the recourse to diagnose or rectify these misalignments. Bridging this divide requires **technical and interaction mechanisms that enable people to adapt—and adapt to—AI representations**.

To do so, **my research re-envisions human-AI alignment as an interactive layer of AI infrastructure that mediates between models and their users** (Figure 1). I develop methods to communicate a model's representations in human-understandable terms, enabling users to build accurate mental models of its behavior [1, 2, 3]. To support this process, I design user-centric alignment metrics that compare model and human domain knowledge, enabling users to determine when to intervene and when to leverage the model's complementary strengths [4, 5, 6]. To operationalize these insights, I create mechanisms that encode users' expectations into the model [7, 8]. Collectively, by giving users agency to understand, measure, and shape their AI alignment, my research advances key human-AI interaction goals, increasing trust, adoption, and collaboration.

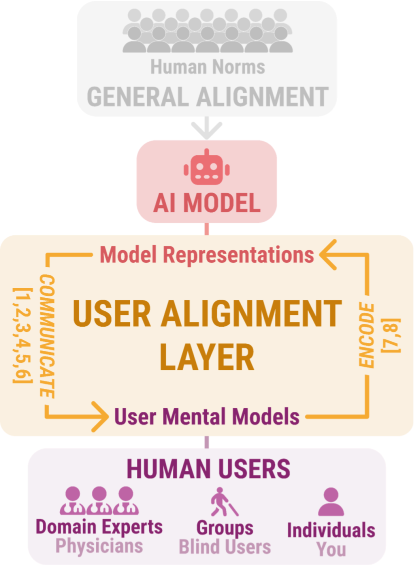


Figure 1. My research empowers users to inspect and improve their AI alignment. To do so, I re-envision alignment as an interactive layer that *communicates* model representations in human-understandable terms and *encodes* users' expectations back into the model.

My research lies at the intersection of human-computer interaction (HCI) and AI, and has resulted in **15 publications** at top conferences in both fields (e.g., CHI, VIS, FAccT, ACL), earning **two Best Paper Honorable Mention awards** and an **Outstanding Paper award**. My research is supported by industry and government funding, and I have won an **MIT PhD Fellowship (~$93K)** and the **Apple Scholars in AIML Fellowship (~$224K)**. Beyond these recognitions, my work has influenced research in diverse domains (e.g., AI architecture [10], accessibility [11], healthcare [12]), design at Apple and IBM, and courses at MIT, Brown, and NYU.

COMMUNICATING MODEL ALIGNMENT TO HUMAN STAKEHOLDERS

Effective human collaboration is predicated on our ability to share and compare mental models. Even in the absence of alignment, people exchange knowledge and reasoning to build trust, coordinate decisions, and outperform their individual capabilities. Yet, human-AI interactions lack this communicative capacity. Alignment is abstracted away from the user, and interfaces reduce interaction to inputs and outputs, forcing people to infer their model's perspective through trial and error. To ensure effective AI collaborations, my research investigates how to communicate the alignment between model representations and users' domain-specific mental models.

Shared Interest: Quantifying Alignment Between Model and User Reasoning

A core challenge in communicating alignment is that humans and models reason over fundamentally different substrates. While humans think in semantic concepts, models explain their decisions through continuous numerical attributions. For example, a dermatologist may diagnose melanoma based on an "irregular border", whereas a model would report importance scores across every pixel channel. Consequently, users (from consultants to AI researchers) are forced to approximate alignment by "eyeballing" individual model explanations, a process that does not scale and is prone to confirmation bias [9].

To facilitate rigorous comparison, I designed *Shared Interest:* a framework of reasoning alignment metrics [4]. By reifying model reasoning via interpretability and human reasoning via annotation, Shared Interest maps these heterogeneous processes into a comparable representation. This transformation enables set-theoretic comparisons that quantify the extent to which a model's reasoning intersects with or diverges from its user.

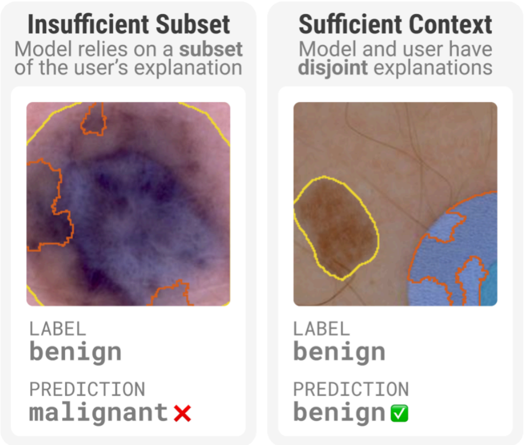


Figure 2. Shared Interest compares model (orange) and human (yellow) explanations, revealing incomplete reasoning (left) and concerning decision-making patterns (right).

Critically, Shared Interest rejects a unidimensional view of alignment. By defining three complementary metrics, it measures not just *if* but *how* reasoning differs. Applying the metrics across models and modalities revealed eight recurring alignment patterns, including reliance on incomplete evidence or additional context (Figure 2). This pattern taxonomy equips users with a vocabulary to systematically audit model decisions, answering previously intractable questions, like "*how often does the model reason like me?*" and "*when are our perspectives complementary?*"

As a result, Shared Interest exposes failure modes that performance metrics obscure. In a clinical case study, a dermatologist used Shared Interest to reveal that a seemingly performant clinical model exploited image artifacts to diagnose melanoma (Figure 2, right). This spurious correlation threatened to waste clinicians' time with incorrect predictions and cause misdiagnoses. By framing misalignment in terms of the dermatologist's thought process, Shared Interest empowered them to intervene by removing artifacts and withholding deployment until the model relied on medical morphology. To scale auditing, I led an MIT master's student in integrating these metrics into a large-scale evaluation pipeline, which automatically surfaced biased reasoning in otherwise indistinguishable models [5].

Abstraction Alignment: Comparing Model and User Domain Knowledge

By revealing differences in reasoning, Shared Interest raises a deeper question: *do models share the domain knowledge that guides users' reasoning?* People organize knowledge into rich semantic graphs, like disease hierarchies and lexical taxonomies, that enable us to generalize logical reasoning. For instance, knowing that "*pneumonia*" is a "*lung infection*" dictates appropriate treatment. Models train on domain corpora, but there is no guarantee that they learn these relationships. Yet, users (particularly experts) expect their AI collaborators to adhere to their domain principles.

To expose whether AI models encode domain knowledge, I developed *Abstraction Alignment* [6] (Figure 3). Moving beyond methods that test for individual human concepts, it evaluates whether models learn human-like conceptual relationships.

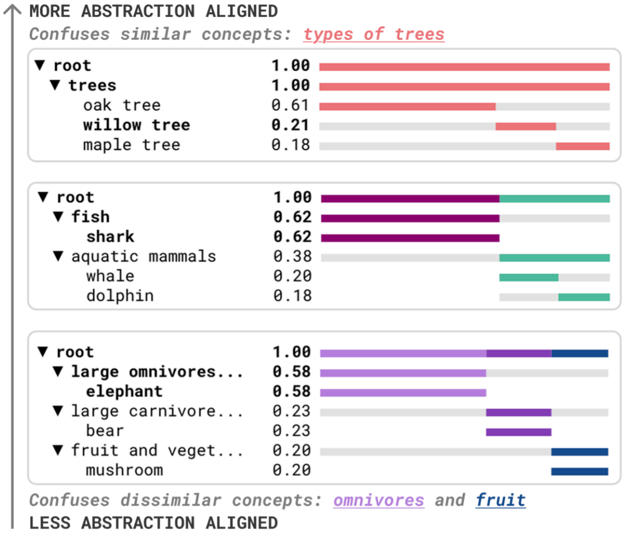


Figure 3. Abstraction Alignment reveals consistency between model behavior and user domain knowledge by measuring the human-similarity of a model's uncertainty.

To do so, it measures the consistency of a model's confusion. If a model understands the domain, it should confuse semantically similar concepts (e.g., mistaking "*pneumonia*" for "*bronchitis*" rather than "*cataracts*"). By mapping the model's output distribution onto the user's knowledge graph, Abstraction Alignment quantifies how closely the model's learned behavior maps to the domain semantics.

As a result, Abstraction Alignment exposes fundamental differences in how models and humans generalize domain knowledge. In evaluations with large language models (LLMs), it revealed a critical behavioral distinction. Unlike humans, who default to correct generalizations when uncertain (e.g., *published in the 2000s*), LLMs prefer specific answers even at the expense of accuracy (e.g., *in June 2018)*. By mapping the model's uncertainty into the user's knowledge graph, Abstraction Alignment inspired LLM researchers to generate novel evaluation metrics and generation strategies that encourage models to generalize to correct abstractions when uncertain, thereby improving alignment.

Abstraction misalignments can also reveal opportunities to refine or expand human knowledge. In a participatory dataset audit, medical experts used Abstraction Alignment to expose discrepancies between how diseases are formally defined (i.e., by the World Health Organization) and how they are diagnosed in practice. Notably, these findings correspond to real-world updates the WHO made to their disease classification system, demonstrating that when latent domain knowledge is made explicit and comparable, alignment tools can serve as mechanisms to improve human understanding.

ENCODING USER EXPECTATIONS INTO MODEL CONSTRAINTS

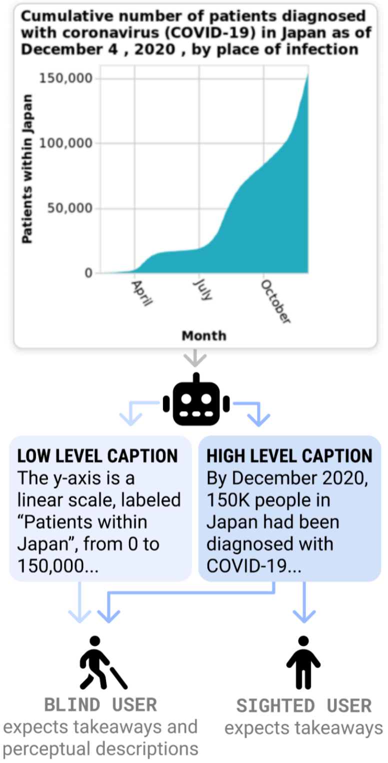
By providing tools to communicate and measure alignment, my work enables users to identify critical domain misalignments and improve their AI collaborations. The natural next challenge is enabling users to express their expectations and bring models into closer alignment with their goals. In doing so, my research bridges the gap between models aligned to general human norms and the contextual, expertise-driven expectations of real users.

VisText: Customizing Model Generation to Diverse Users

I investigated user-specific alignment in the context of AI-automated chart captioning. While captions improve data literacy, studies reveal substantial variation in the caption content that different audiences expect. Lay readers prefer high-level takeaways, whereas blind users value perceptual details. In this domain, alignment requires models to adapt caption generation to their users.

In response, I developed *VisText* [7], a framework for customizing models to users' expectations (Figure 4). Monolithic captions inevitably misrepresent some users, so VisText solves this problem by restructuring the underlying data representation. In particular, the VisText corpus decomposes chart captions into semantic levels, ranging from low-level visual descriptions to high-level insights.

By fine-tuning vision-language transformers on this stratified data, models dynamically adjust their caption detail, producing perceptual descriptions for blind users and analytical summaries for sighted readers. As a result, VisText demonstrates that better representations of user intent produce models that better reflect user needs. VisText has informed accessibility at Apple and been adopted in standard LLM benchmarks, improving real-world human-AI experiences.


Cumulative number of patients diagnosed with coronavirus (COVID-19) in Japan as of December 4 , 2020 , by place of infection
Patients within Japan
150,000
100,000
50,000
0
April
July
October
Month
LOW LEVEL CAPTION
HIGH LEVEL CAPTION
BLIND USER
expects takeaways and perceptual descriptions
SIGHTED USER
expects takeaways

The y-axis is a linear scale, labeled "Patients within Japan", from 0 to 150,000...

By December 2020, 150K people in Japan had been diagnosed with COVID-19...

Figure 4. VisText's multimodal chart captioning models customize their output per user, including takeaways for sighted readers and perceptual content for blind users.

FUTURE RESEARCH AGENDA

I aim to design a future where AI models empower people, enabling them to interrogate and reshape models on demand to improve trust, collaboration, and knowledge discovery.

Translating User Intent into Reliable Model Behavior

Despite increasingly capable LLMs, people struggle to elicit expected responses. The rise of "prompt engineering"—esoteric heuristics ranging from "*provide step-by-step instructions*" to "*be rude*"—exposes a fundamental representation mismatch. Although natural language mimics human conversation, people's mental models of communication differ significantly from LLM interpretation. However, my research is posed to address this, designing interfaces that translate between user intent and model execution. For instance, "prompt linters" could flag ambiguous phrasing and suggest structural refinements that shift a prompt closer to the user's goal. Additionally, declarative syntax could augment natural language prompts, allowing users to specify explicit constraints (e.g., `output[`*"follow the Google style guide"*`]` or `RAG[`*"only peer-reviewed citations"*`]`). These constraints could be compiled into verifiable functions that guarantee the model's output satisfies user intent before being returned. By identifying prompts that fail verification, the interface generates "hard-negative" preference data to increase model alignment. This creates a positive feedback loop: as models better adhere to user constraints, the interface can evolve to support more sophisticated specifications, increasing user control alongside model capability.

The Value of Human-AI *Misalignment*

Current alignment methods, like RLHF and Constitutional AI, aim for maximal alignment. While this is appropriate for general safety, it creates cognitive redundancy in human-AI collaborations—i.e., if a model perfectly mirrors its user, it cannot expand their thinking. Instead, my future research explores *strategic misalignment*. With my user-centric alignment metrics [4, 5, 6], we can quantify the representational distance between a user's mental model and a candidate pool of AI collaborators. This makes it possible to algorithmically select, or even train, an optimally misaligned model—one that shares the user's core knowledge but offers an orthogonal reasoning process. Much like in diverse human teams, these models would preserve common ground while introducing complementary perspectives that improve collective performance. By precisely modulating the degree of human-AI alignment, empirical studies could identify the ideal representational distance between collaborators and generate theories that generalize back to the design of effective human teams. Ultimately, this line of research aims to ensure that AI collaborators upskill, rather than replace, their users.

Alignment as a Method for Co-evolving with Superintelligent AI

As AI models trend towards superhuman performance, concerns grow that they will replace human intelligence. However, by communicating alignment to users [1, 2, 4, 6], my research suggests that AI can be a catalyst for human learning. History shows that novel representations (e.g., the Cartesian plane) fundamentally expand human reasoning. As models surpass human performance, they will likely develop new, superior representations. While my prior work mapped model representations to existing human concepts [4,6], my long-term agenda focuses on translating *novel* model representations into human-*learnable* concepts. Realizing this vision requires developing interpretability methods to identify latent patterns, designing interfaces that synthesize these patterns into coherent human abstractions, and conducting user studies to validate their pedagogical compatibility. By exploring this framework in domains where models already outperform people (e.g., games, protein folding) and where users have something to learn (e.g., students), I plan to study how people can acquire model-generated representations. In doing so, I aim to ensure that as AI becomes more capable, so do people, allowing us to "*see further by standing on the shoulders of [AI] giants.*"

References

Asterisks (*) denote equal contribution.

[1] Embedding Comparator: Visualizing Differences in Global Structure and Local Neighborhoods. **Angie Boggust***, Brandon Carter*, and Arvind Satyanarayan. *ACM Intelligent User Interfaces (IUI) 2022.* 🏅 **Best Paper Honorable Mention Award**

[2] Compress and Compare: Interactively Evaluating Efficiency and Behavior Across ML Model Compression Experiments. **Angie Boggust***, Venkatesh Sivaraman*, Yannick Assogba, Donghao Ren, Dominik Moritz, and Fred Hohman. *IEEE Transactions on Visualization & Computer Graphics (VIS) 2024.*

[3] Semantic Regexes: Auto-Interpreting LLM Features with a Structured Language. **Angie Boggust**, Donghao Ren, Yannick Assogba, Dominik Moritz, Arvind Satyanarayan, and Fred Hohman. *Under Review at the International Conference on Learning Representations (ICLR) 2026.*

[4] Shared Interest: Measuring Human-AI Alignment to Identify Recurring Patterns in Model Behavior. **Angie Boggust**, Benjamin Hoover, Arvind Satyanarayan, and Hendrik Strobelt. *ACM Human Factors in Computing Systems (CHI) 2022.* 🏅 **Best Paper Honorable Mention Award**

[5] Explanation Alignment: Quantifying the Correctness of Model Reasoning At Scale. Hyemin Bang, **Angie Boggust**, and Arvind Satyanarayan. *Explainable Computer Vision (eXCV) Workshop at the European Conference on Computer Vision (ECCV) 2024*.

[6] Abstraction Alignment: Comparing Model-Learned and Human-Encoded Conceptual Relationships. **Angie Boggust**, Hyemin Bang, Hendrik Strobelt, and Arvind Satyanarayan. *ACM Human Factors in Computing Systems (CHI) 2025*.

[7] VisText: A Benchmark for Semantically Rich Chart Captioning. Benny J. Tang*, **Angie Boggust***, and Arvind Satyanarayan. *The Annual Meeting of the Association for Computational Linguistics (ACL) 2023.* 🏅 **Outstanding Paper Award**

[8] DiffusionWorldViewer: Exposing and Broadening the Worldview Reflected by Generative Text-to-Image Models. Zoe De Simone, **Angie Boggust**, Arvind Satyanarayan, and Ashia Wilson. *arXiv 2024.*

[9] Saliency Cards: A Framework to Characterize and Compare Saliency Methods. **Angie Boggust***, Harini Suresh*, Hendrik Strobelt, John Guttag, and Arvind Satyanarayan. *ACM Conference on Fairness, Accountability, and Transparency (FAccT) 2023*.

[10] One Wide Feedforward is All You Need. Pires et al. Empirical Methods in Natural Language Processing (*EMNLP) 2023*

[11] Chart4Blind: An Intelligent Interface for Chart Accessibility Conversion. Moured et al. *ACM Intelligent User Interfaces (IUI) 2024.*

[12] Inconsistency between Human Observation and Deep Learning Models: Assessing Validity of Postmortem Computed Tomography Diagnosis of Drowning. Zeng et al. *Journal of Imaging Informatics in Medicine 2024.*