# Khanh Nguyen Research Statement

The field of AI has long pursued the goal of creating autonomous agents. While this focus has impressively advanced AI agents' ability to solve tasks independently, it is insufficient to turn them into helpful assistants of humans. Agents designed solely for automation *lack communication with humans while performing tasks*, which renders them potentially unsafe and ineffective when faced with new environmental conditions.

**I develop AI agents that communicate with humans to assist them more safely and effectively.** These agents extend beyond traditional autonomous agents. They make decisions independently on tasks they have mastered, yet in unfamiliar situations, they proactively inform human operators of their uncertainties and seek guidance. Even when they cannot solve a problem, they have the incentive and skills to help human collaborators solve it. They can also learn from diverse types of human feedback.

To build such agents, I invented novel algorithmic frameworks that extend traditional learning frameworks with *human-inspired cognition and communication models*. My work addresses the fundamental problems in human-AI communication, specifically:

- **The *listening* problem: learning from human feedback (§1)**. I made early contributions to the development of reinforcement learning from human feedback (RLHF) methods for text generation [2]. A slight variant of the algorithm I proposed remains the state-of-the-art approach for fine-tuning large language models (LLMs), offering performance comparable to OpenAI's approach while being simpler to implement and more memory- and compute-efficient. Toward human-like learning, I pioneered frameworks for learning from language feedback with theoretical guarantees [5, 11].
- **The *speaking* problem: learning to convey uncertainties and instruct humans (§2)**. My work on calibration [1] inspired uncertainty assessments of modern neural networks and approaches to out-of-distribution detection. Enabling agents to convey richer information, I spearheaded the development of AI agents that can ask for help [3, 4, 6, 8] and use language to assist humans in navigation [7, 9, 10].

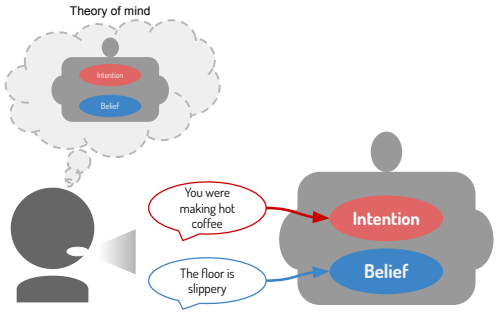
More broadly, my research sits at the intersection of sequential decision-making (reinforcement and imitation learning), natural language processing, and AI safety. It has resulted in publications at top-tier conferences of multiple AI subfields (ICML, EMNLP, ACL, CVPR).

## 1 Learning from Human Feedback

**Reinforcement learning from rating feedback.** Large language models (LLMs) have revolutionized how humans interact with and benefit from AI systems. Reinforcement learning from human feedback (RLHF) is a crucial step in the training procedure of today's state-of-the-art LLMs. This approach employs a reinforcement learning (RL) algorithm to improve a model using human ratings (or rankings) of its behaviors. Early work on RLHF either concerned non-neural-network models [15] or experimented with ratings that are clean and provided even for partial outputs [14, 16], which were favorable for RL algorithms and under-represented the challenges of real-world settings. **In 2017, I demonstrated for the first time the feasibility of using only noisy, complete-output ratings to improve the performance of a neural text generator [2]**. This work was followed by studies that used real human ratings at eBay [17] and OpenAI [18, 19]. These developments led to the advent of InstructGPT [20] which popularized RLHF. The training recipe that I proposed in [2] directly inspired the RLOO algorithm [21], which has been shown to *outperform OpenAI's PPO approach in fine-tuning modern LLMs in terms of speed and memory efficiency without compromising output quality*. This algorithm has been integrated into the popular *HuggingFace TRL library* (9.9K stars on Github). In addition to showcasing the potentials of RLHF, I also illustrated the degradation of this method due to various imperfections in human ratings. This issue continues to be relevant in contemporary expert discussions on the drawbacks of RLHF [22].

**Learning from language feedback.** Rating feedback is overly restricted for humans to express intentions. Language is a more natural and efficient medium for this purpose. Decades of socio-cognitive research have revealed that humans use language primarily to *influence the cognitive processes they attribute to others*. This finding

to me entails that, to support language-based communication, AI systems must enable humans to (1) build accurate theory-of-mind models of their cognitive processes and (2) use language to manipulate those processes. I operationalized this insight through the development of two frameworks: Interactive Learning from Activity Description (ILIAD) and Language-Guided World Model (LWM). These frameworks construct agents with *language-parameterized* mental representations, making it possible for humans to interpret and adapt their cognitive processes through verbal communication. Furthermore, these mental representations are modeled by well-defined mathematical objects in sequential decision-making theory, facilitating theoretical analyses.




ILIAD [5] deals with feedback describing the *intention* of an agent. Suppose an agent aims to "make iced coffee" but forgets to put ice, a human can say "you were *making hot coffee*" to correct its intention. The agent collects such feedback on multiple tasks to learn a probabilistic mapping from intention to behavior. At test time, the human can tell the agent which intention they want it to fulfill. In the paper, we proved the convergence of ILIAD in a bandit learning setting. To the best of my knowledge, **ILIAD is the first language-based learning framework to provide such an asymptotic convergence guarantee**. Although recent advances in LLMs give rise to agents that can "improve" with language feedback, this approach lacks guaranteed convergence, as the ways these models adapt to language feedback remains theoretically poorly understood.

Meanwhile, LWM [11] extends model-based RL to allow the incorporation of language feedback that targets *beliefs* about an environment. Similar capabilities significantly reduce human communication effort in learning and teaching. For example, a person can simply say "the floor is slippery!" to cause another person to handle *every* object in a room with greater caution. LWM is a theoretically grounded approach; the sub-optimality of the agent performance in this approach can be bounded using tools borrowed from model-based RL. My paper showcased the potential of LWM in enhancing AI safety by simulating a scenario in which an agent uses its LWM to *generate and discuss visual plans with a human* instead of acting heedlessly with imperfect information. We also proposed a robust architecture for LWM that can generalize to *compositionally novel* feedback.

## 2 Learning to convey uncertainties and instruct humans

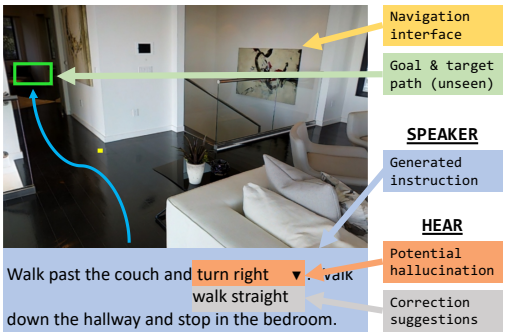
**Conveying uncertainty through calibrated probabilities.** Calibration guarantees that the probabilities estimated by a model accurately reflects its empirical chance of being correct. **My paper [1] introduced calibration analysis for structured outputs (e.g., sequences, graphs).** The paper inspired methodologies for measuring and visualizing the calibration of modern neural networks [25, 27, 30, 31] and theoretical frameworks for calibration [28, 33]. It has been frequently referenced in prominent papers in AI evaluation and safety (e.g., BigBench [23], a large-scale AI evaluation suite constructed by 450 authors, and Hendrycks et al. [24], a foundational paper in AI safety). In particular, my paper demonstrated an unsupervised technique for identifying data points that are difficult for a model by examining output probabilities. This technique motivated a well-known baseline for out-of-distribution detection ([26], ∼3,600 citations), sparking a productive line of research on this problem. We recently applied calibration analysis to fine-tuned LLMs [12], revealing that these models are over-confident but their output probabilities remain strongly predictive of their correctness.

**From uncertainty to uncertainties: learning to ask for help.** Although probabilistic calibration improve trustworthiness, effective collaboration with humans requires the ability to share *richer and more insightful uncertainty information*. This motivated me to develop agents capable of asking for help. This capability improves safety by alerting human users to potential risks. It can also dramatically boost task performance with minimal human effort, if agents can intelligently decide when and what to ask.

To facilitate empirical research, I created the VLNA and HANNA human-assisted navigation tasks. These tasks require a virtual agent to complete object-finding tasks in photo-realistic simulations of houses. I build simulated humans that, upon request, advise the agent on the next actions to take. HANNA additionally requires the agent to interpret *language advice*. I proposed a novel imitation learning approach, which synthetically constructs progress labels to teach agents to *predict whether they will make progress from a current state*. This approach lifted performance of an autonomous agent by almost *six times* and outperformed the ask-every-k-steps and ask-at-random-steps strategies while making less help requests. **VLNA and HANNA have effectively promoted research on robots that ask for help, inspiring many variants of this problem [34, 35, 36].**


Toward learning more human-like behavior, I proposed two frameworks: HARI and CEIL. HARI [6] is a general hierarchical RL framework that addresses the problems of when and what to ask jointly. Notably, it allows an agent to ask questions to decompose a problem into simpler subproblems (e.g.,"should I go to the kitchen to get a mug?" The framework trains agents to *ask questions that are truly helpful to their decision making*. This approach contrasts with imitation learning approaches that force agents to reproduce human-generated questions without considering the utility of those questions to the agents. Meanwhile, CEIL [8] teaches agents to *ask increasingly more abstract questions over time*, increasing the communication efficiency between the student and the teacher over time. This phenomenon is characteristic of human communication but is absent in frameworks like imitation and reinforcement learning, which employ static shared "languages" (label or reward). We successfully induced progressively efficient communication in a 2D MineCraft problem, in which our agent initially verified micro-decisions (e.g., asking "should I go left?") and gradually transitioned to confirming high-level intentions (e.g., asking "should I get coal?").

**Generating navigation instructions.** In this line of work, I develop systems that generate language instructions to aid human navigation in residential buildings. Such systems can find many useful applications: helping residents find missing objects, guiding visitors or maintenance workers, assisting in search and rescue missions, aiding people with visual or cognitive impairments. In [9], I and my student designed a cognitive test that exposed the deficiency of the pragmatic-reasoning capability of neural vision-language models. We introduced an ensemble method to construct a robust simulation of human listeners. Incorporating this module increased the success rate of guiding *real* humans in simulated houses by 11%. **This paper received an outstanding paper award at the ICML'23 workshop on theory of mind.** In follow-up work [7, 10], dealing with the inevitable failure of vision-language models in unfamiliar situations, we designed an assistant system that could *guide humans to navigate successfully despite generating fallible instructions*. To achieve this, we introduced *evaluative models* that could detect errors in an instruction and compose a list of potential corrections. Without changing the generated instructions, simply highlighting errors and offering correction suggestions boosted the human navigation success rate by 13%. **Our paper is one of the first to demonstrate that effective communication of uncertainty can enhance human decision-making.** This finding opens an entire new dimension for model development.




## 3 Future plans

The arrival of LLMs may give the impression that human-like communication is nearly achieved. However, I argue that these models lack the cognitive infrastructure needed to support the most effective and efficient forms of human communication. The black-box nature of LLMs makes it difficult for (non-expert) humans to construct

mental models of their cognitive processes and reason strategically to manipulate those processes. Their opacity also hinders the development of theoretically grounded approaches. Furthermore, LLMs are currently trained to pursue improper communication goals, which drive them toward imitative or sycophantic behaviors rather than fulfilling the assistance goal of helping their human owners succeed in the real world.

Toward principled, effective human-like communication, my long-term goal is to equip AI agents with more adequate cognitive infrastructure and communication goals. More specifically, I plan to:

1. Build *cognitive models* that enhance not only AI agents' reasoning capabilities but also human capabilities of reasoning about and controlling their thinking processes;
2. Formulate *intrinsic motivations* that engender efficient and beneficial communication behaviors.
3. Develop technologies to construct practical *world and human simulations*, which are essential to make empirical research on human-AI communication scalable, safe, and reproducible.

My past research has provided useful principles and insights to address these challenges. Below, I outline several short-term research agendas aimed at demonstrating the effectiveness of those approaches through impactful applications, and pushing the boundaries of their capabilities.

**Alignment with humans who can be wrong about the world.** In an unpublished work [13], I propose a theoretical framework to characterize the drawbacks of RLHF when used for alignment with humans who have misconceptions about the world. Notably, I show that this framework encourages AI agents to *manipulate* humans beliefs to prove its effectiveness. I remedy this problem by defining a new alignment objective that evaluates the behavior of an agent based on its effectiveness in *real* world rather than in a human's imaginary world. This objective consequently motivates an agent not only to learn, but also to *truthfully teach* humans about the world. This brings to attention a set of teaching problems that are orthogonal to learning from human feedback. My future plan is to develop robust teaching algorithms that involve Bayesian inference methods to detect false beliefs of humans and pragmatic language-generation techniques to produce accurate and relevant world descriptions.

**Learning to yield and request control.** I and colleagues are building a benchmark for the problem of deciding when a robot should signal a human to take or yield control of it. This problem extends the when-to-ask problem, as the robot has a new choice of asking the human to *stop* helping it. Making this choice wisely could further reduce human assistance effort. Our goal is to incorporate diverse environments and human models, and provide clean implementations of popular approaches to provide an off-the-shelf toolkit and facilitate the development of robust techniques. We will be releasing the first version of the benchmark in a couple of months, which features video-game and robotics environments, and several families of approaches (probability-based, out-of-distribution detection, RL). The benchmark will provide the necessary first step toward solving this problem.

**Training LLMs with language feedback that teaches them how to think.** LLMs forge their capabilities by observing myriad traces of human linguistic behaviors. These traces, however, present only a narrow glimpse of human thinking processes. This approach therefore results in mediocre thinking capabilities. A fully supervised learning approach that collects thoughts from humans would be too expensive. I put forward two ideas to reduce this cost: (i) fine-tune LLMs to generate language feedback (ii) synthetically construct language feedback by collecting a cheaper form of feedback (e.g., collecting preferences to teach agents how to improve solutions). The language feedback can be incorporated through an ILIAD-like process, which facilitates theoretical guarantees.

**Software-using AI designer agents.** The goal is to create agents capable of using human-developed software to build complex world simulations. Such agents could dramatically boost the productivity of human designers. In research, they can be employed to automatically design environments for testing AI agents, offering improved comprehensiveness and realism. They can also serve as theory-of-world components of general-purpose agents, facilitating robust planning. Such agents also enable humans to modify their "world model" by directly changing the simulation code. This approach can be effective when high-precision control is demanded. I plan to construct a benchmark for this problem featuring simple environment-building packages like MiniGrid and AI-Thor, and gradually transition more complicated software like PyGame or Unity.

# References

## Authored papers


[1] **Nguyen**, O'Connor. *Posterior calibration and exploratory analysis for natural language processing models*. EMNLP 2015.

[2] **Nguyen**, Daumé III, Boyd-Graber. *Reinforcement learning for bandit neural machine translation with simulated human feedback*. EMNLP 2017.

[3] **Nguyen**, Dey, Brockett, Dolan. *Vision-based navigation with language-based assistance via imitation learning with indirect intervention*. CVPR 2019.

[4] **Nguyen**, Daumé III. *Help, Anna! Visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning*. EMNLP 2019.

[5] **Nguyen**, Misra, Schapire, Dudik, Shafto. *Interactive learning from activity description*. ICML 2021.

[6] **Nguyen**, Bisk, Daumé III. *A framework for learning to request rich and contextually useful information from humans*. ICML 2022.

[7] Zhao, **Nguyen**, Daumé III. *Hallucination detection for grounded instruction generation*. EMNLP 2023 Findings.

[8] Zheng, **Nguyen**, Daumé III, Huang, Narasimham. *Progressively efficient learning*. Workshop on Intrinsically Motivated Open-ended Learning (NeurIPS 2023).

[9] Zhao, **Nguyen**, Daumé III. *Define, evaluate, and improve task-oriented cognitive capabilities for instruction generation models*. ACL 2023 Findings & **Outstanding Paper Award** at Workshop on Workshop on Theory of Mind in Communicating Agents (ICML 2023).

[10] Zhao, **Nguyen**, Daumé III. *Successfully guiding humans with imperfect instructions by highlighting potential errors and suggesting corrections*. EMNLP 2024.

[11] Zhang, **Nguyen**, Tuyls, Lin, Narasimhan. *Language-guided world models: A model-based approach to AI control*. Workshop on Spatial Language Understanding and Grounded Communication for Robotics (ACL 2024).

[12] Plaut, **Nguyen**, and Trinh. *Probabilities of chat LLMs are miscalibrated but still predict correctness on multiple-choice Q&A*. Under Review.

[13] Trinh, **Nguyen**. *Alignment with humans who can be wrong about the world*. Under Review.


## Reinforcement Learning from Human Feedback


[14] Ranzato et al.. *Sequence Level Training with Recurrent Neural Networks*. ICLR 2016.

[15] Sokolov et al.. *Bandit Structured Prediction for Learning from Partial Feedback in Statistical Machine Translation*. MT Summit XV.

[16] Bahdanau et al.. *An Actor-Critic Algorithm for Sequence Prediction*. ICLR 2017.

[17] Kreutzer et al.. *Can Neural Machine Translation be Improved with User Feedback?*. NAACL-HLT 2018 (industry track).

[18] Stiennon et al.. *Learning to summarize from human feedback*. NeurIPS 2020.

[19] Ziegler et al.. *Fine-Tuning Language Models from Human Preferences*. ArXiv 2019.

[20] Ouyang et al.. *Training language models to follow instructions with human feedback*. NeurIPS 2022.

[21] Ahmadian et al.. *Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs*. ArXiv 2024.

[22] Casper et al.. *Open problems and fundamental limitations of reinforcement learning from human feedback*. ArXiv 2023.


## Calibration


[23] Srivastava et al.. *Beyond the imitation game: Quantifying and extrapolating the capabilities of language models*. ArXiv 2022.

[24] Hendrycks et al.. *Unsolved problems in ML safety*. ArXiv 2021.

[25] Hendrycks et al.. *Calibration of Encoder Decoder Models for Neural Machine Translation*. ArXiv 2019.

[26] Hendrycks et al.. *A baseline for detecting misclassified and out-of-distribution examples in neural networks*. ICLR 2017.

[27] Nixon et al.. *A baseline for detecting misclassified and out-of-distribution examples in neural networks*. ICLR 2017.

[28] Kumar et al.. *Verified uncertainty calibration*. NeurIPS 2019.

[29] Kuleshov et al.. *Accurate Uncertainties for Deep Learning Using Calibrated Regression*. ICML 2018.

[30] Ott et al.. *Analyzing Uncertainty in Neural Machine Translation*. ICML 2018.

[31] Desai et al.. *Calibration of Pre-trained Transformers*. EMNLP 2020.

[32] Hendrycks et al.. *Augmix: A simple data processing method to improve robustness and uncertainty*. ICLR 2020.

[33] Kuleshov et al.. *Calibrated structured predictio*. NeurIPS 2015.


## Vision-Language Navigation


[34] Anderson et al.. *Vision-and-dialog navigation*. CoRL 2019.

[35] Ren et al.. *Robots that ask for help: Uncertainty alignment for large language model planners*. CoRL 2023.

[36] Padmakumar et al.. *Teach: Task-driven embodied agents that chat*. AAAI 2022.