Research Statement
Kaiyu Yang

As a researcher in artificial intelligence and machine learning, my long-term goal is to build effective, robust, and interpretable machine learning that is trustworthy for deployment in real-world applications and making consequential decisions about people. During my Ph.D. studies, I have approached this goal from three perspectives:

(1) **Improving the fundamental capabilities**: Machine learning has made significant progress on System-1 capabilities—intelligent behaviors that are fast, automatic, and unconscious—such as recognizing objects, understanding simple sentences, and driving cars in clear road conditions. However, it still struggles with System-2 capabilities, which typically involve sequential, logical, and conscious reasoning, such as solving math equations and writing computer programs. To bridge the gap, my research aims to develop machines that reason—in ways that are precise, systematic, interpretable, and robust to ambiguity in real-world environments.

(2) **Rigorously gauging the performance**: Before deploying a machine learning model in the real world, we must thoroughly understand its performance and failure modes. The common practice is to use metrics on benchmark datasets as proxies. However, researchers have repeatedly found that models can perform superficially well on specific benchmarks but generalize poorly. This is partially attributed to dataset bias; the model exploits spurious correlations in the data to learn shortcuts rather than the correct solution [5]. For example, it may label an image as “duck” whenever it sees grass because most ducks are on the grass in the training data. My research enables more rigorous evaluation of the performance of machine learning models through novel crowdsourcing methods for reducing dataset bias.

(3) **Addressing societal concerns such as fairness and privacy**: Machine learning models are being deployed to make critical decisions about people, e.g., hiring, loan approval, and college admission. It has raised concerns about whether the decisions are fair or whether the decision-making process respects privacy. Many concerns are rooted in datasets: Historical data inevitably contains biases against underrepresented groups. And existing data collection practices frequently overlook fairness and privacy. Therefore, I address these concerns by intervening at the dataset level. By analyzing and fixing existing datasets, my research gains insights that inform future dataset creation efforts.

The three tenets of my research—improving the fundamental capabilities of machine learning models, rigorously gauging their performance, and addressing their societal concerns—collectively contribute to a future of machine learning that we can trust, even when the stakes are high. I will elaborate more on each of them and discuss future directions.

1. **Neuro-symbolic Reasoning**

   ![Figure 1: A progressive spectrum of reasoning, from formal logic to natural language.](image)

---

1 “System 1” and “System 2” are terms popularized by Daniel Kahneman’s book “Thinking, Fast and Slow.”
Reasoning is a core System-2 capability that machines still struggle with. I develop models that represent reasoning via symbolic proofs by combining the complementary strengths of machine learning and symbolic reasoning. Symbolic reasoning is precise and generalizes systematically to unseen scenarios. But it has been restricted to domains amenable to rigid formalization. In contrast, machine learning is flexible enough to handle noisy and ambiguous domains. But predominant models, such as deep neural networks, are notoriously uninterpretable, data-hungry, and incapable of generalizing outside the training data distribution. Integrating the strengths of both approaches is essential for building flexible reasoning machines with precise and systematic generalization. However, due to the discrete nature of symbolic reasoning, such integration may require a radical departure from the predominant paradigm of gradient-based learning. My research tries to answer what that alternative form of learning might look like.

Concretely, my work spans a progressive spectrum of reasoning (Fig. 1), including reasoning in formal domains such as automated theorem proving [11], in a “quasi-natural” language compatible with natural language but consisting of symbolic rules [12], as well as in free-form natural language.

**Reasoning in formal logic** [11]. Automated theorem proving is a classical reasoning task in AI, which is critical to building trustworthy systems such as verified software/hardware [10, 9, 8]. It aims to find proofs automatically for theorems in formal logic. Classical methods represent theorems at a low level (e.g., conjunctive normal form in first-order logic) and focus on searching for proofs efficiently in a large space with hand-crafted heuristics. More recent methods also learn search heuristics via machine learning. Despite substantial progress, they are still incapable of scaling to large formal theories of practical significance.

Formal reasoning in large theories requires manual efforts from human experts [7]. In a software environment called proof assistant, humans define mathematical objects formally and prove theorems by entering a sequence of commands called tactics. The tactics capture high-level proof techniques such as induction, leaving low-level details to the proof assistant (Fig. 1 Left). However, this approach is notoriously labor-intensive and requires extensive training of human experts.

My work uses machine learning to prove theorems efficiently by automating the interaction with proof assistants. Our key insight is to perform automated reasoning at a higher level of abstraction than first-order logic—Proof assistants offer such a high-level formalism that resembles human mathematical reasoning. Compared to classical provers, our approach benefits from not only high-level reasoning but also the unique opportunity of learning from human-written proofs. I introduce CoqGym [11]—a large-scale dataset and learning environment containing 71K human-written proofs in the Coq proof assistant. CoqGym is one of the first works on learning to prove theorems leveraging proof assistants. It enables training machine learning models that can prove theorems not previously provable by automated methods, and it has spurred subsequent research along this direction.

**Reasoning in “quasi-natural” language** [12]. Beyond formal reasoning, human reasoning is far more general and flexible, capable of handling domains involving natural language and commonsense knowledge that are infeasible to formalize. In these domains, classical AI has attempted symbolic reasoning with manually constructed rules, but had limited success and tended to produce narrow and brittle systems. Current leading methods are almost exclusively neural networks. However, they require a large amount of training data and can suffer from poor generalization, e.g., to longer reasoning chains unseen in training. More importantly, neural networks are black boxes that are hard to interpret and verify. Such lack of interpretability is undesirable in high-stake applications.

My research [12] challenges the conventional wisdom that neural networks are far superior to rule-based systems. I hypothesize that rule-based systems historically underperformed due to the difficulty in manually constructing rules. Therefore, learning rules from data could be key to building effective rule-based systems, but it may require a different kind of learning than gradient descent.

Specifically, I investigate how to learn a rule-based system that reasons symbolically but is flexible enough to work with natural language and other domains difficult to formalize. As an initial solution, I introduce MetaQNL, a “quasi-natural” language that can express both formal logic and natural language (Fig. 1 Middle), and MetaInduce, a learning algorithm that induces MetaQNL rules from training data leveraging existing MAX-SAT solvers. This approach is radically different from gradient-based learning, but experiments have shown encouraging results. Our purely symbolic
approach requires less training data, learns much more compact models, and achieves state-of-the-art accuracies on multiple reasoning benchmarks. It produces not only answers but also symbolic proofs that are both interpretable to humans and checkable by computers. Moreover, our method can handle noise and ambiguity, which have been known to be challenging for rule-based systems.

**Reasoning in natural language.** Free-form natural language presents unique challenges for reasoning. First, it requires the model to understand the syntax and semantics of natural language, which is incredibly rich and complex. Second, reasoning in natural language is often fuzzy and ambiguous, without the same degree of rigor as symbolic reasoning. To address these challenges, I believe the best tool currently available is large pretrained language models such as BERT [4]. Pretrained on huge amounts of texts, these models learn to capture syntax and semantics in continuous vector space. Further, they excel at handling fuzzy and ambiguous inputs. My ongoing work explores building powerful and flexible reasoning systems for free-form natural language by integrating symbolic reasoning structures with large pretrained language models.

2 Mitigating Bias in Benchmark Datasets

Machine learning datasets often contain bias that allows the model to perform superficially well without solving the underlying task. To gauge the performance rigorously, my research has developed novel crowdsourcing methods for reducing dataset bias.

**Adversarial crowdsourcing [14].** In adversarial crowdsourcing, I explicitly ask crowd workers to construct challenging data examples. I use a model trained on available data to make predictions on new examples from the worker. And the worker succeeds only if they construct an example to make the model fail. Datasets constructed through adversarial crowdsourcing contains significantly less bias and more interesting examples in the long tail.

**Minimally contrastive data collection [6].** In minimally contrastive data collection, we collect pairs of almost identical examples with different labels. As a result, the dataset has minimal spurious cues for the model to exploit, leading to not only more rigorous evaluation but also better sample efficiency when used for training.

The study of dataset bias relates naturally to systematic generalization and has led me to reflect on the limitations of training/evaluating a generic neural network on large datasets. My work on reducing dataset bias approaches the limitation from the evaluation perspective, but it does not touch upon developing models that generalize systematically rather than simply exploiting spurious statistical patterns. I believe the solution to the latter problem requires a more principled modeling approach, e.g., the approach I am actively exploring in my reasoning work.

3 Socially Responsible Machine Learning

As machine learning advances from research lab curiosities into real-world applications making consequential decisions in people’s daily lives, it is imperative to address the potential societal consequences. My research has focused on fairness and privacy. Many issues are rooted in the datasets we use to develop machine learning models. Therefore, I reveal and mitigate issues in arguably one of the most influential datasets in machine learning—ImageNet [3].

**Fairness [13].** Biases and prejudices in human society have inevitably made their way into our data collection process and the datasets for training machine learning models. For example, in one of the first benchmark datasets—TIMIT—more than 90% of the subjects are white [1]. My work conducts the first study on the demographic distribution in ImageNet. By carefully annotating the gender, age, and skin color of people in the images, I highlight the lack of representation of certain demographic groups in ImageNet, and propose first steps towards mitigation.

**Privacy [15].** ImageNet is not a people-centric dataset, but it contains images of incidental people, whose privacy becomes a concern as the dataset is widely used. My research mitigates ImageNet’s privacy issue by annotating faces in the images and obfuscating the faces to construct a privacy-enhanced version of the dataset. Extensive experiments demonstrate that face obfuscation does not compromise ImageNet’s utility as an image classification benchmark and a pretraining dataset.

Despite focusing on ImageNet, the approach and insights in this line of research are applicable to a broader range of existing large-scale datasets that have fueled the recent progress in machine
learning. More importantly, my research on existing datasets informs future data collection practices that put fairness and privacy among the first design principles.

4 Future Directions

My work has scratched the surface of robust, interpretable, and trustworthy machine learning in my long-term vision. In the future, I would like to further pursue this goal in the following directions:

**Proper evaluation of reasoning.** Reasoning is the North Star that endows machines with a wide range of System-2 capabilities. But to get there, we need concrete milestones—suitable benchmarks for evaluating our progress on machine reasoning. Existing benchmarks are insufficient; they are either too small, synthetic, or biased, leading to saturating performance of existing methods that simply train a large and generic neural network with little reasoning capabilities.

Building on my experience in rigorously evaluating machine learning models, I plan to construct benchmarks for evaluating reasoning properly. I believe such benchmarks must encompass three components missing in conventional benchmarks but critical to reasoning: compositionality, zero-shot testing, and out-of-distribution testing. Reasoning is inherently compositional, and it requires systematic generalization to novel scenarios. In contrast, current machine learning models struggle with these abilities. This line of research would enable the machine reasoning community to have a clear target and make rapid progress.

**Compositional generalization in large language models.** Large pretrained language models such as GPT-3 [2] have demonstrated superior performance on many NLP tasks. However, as the model and data sizes continue to grow, they still fail to learn compositionality. People have found prominent failing modes: These models are weak at arithmetics and sentences with complex logic. And they can be easily confused by inserting “not” in the input sentence. These failing modes show that the models cannot handle novel examples composed of existing primitives in ways unseen in training. Due to the nature of compositionality, such novel examples always exist, and it is impossible to cover all of them even with larger and larger training data.

In the future, I aim to thoroughly understand the behaviors of large language models on compositional examples. Furthermore, I will explore inductive biases for these models to learn compositionality more effectively. This line of research will not only fix the failing modes but also make language models more powerful and widely applicable, e.g., to downstream tasks that require reasoning.

**Multidisciplinary applications of reasoning.** My existing work focuses on solving the foundation of reasoning, but we have seen emerging applications in various domains, including chemistry, biology, and law. In chemistry, for example, reactions chain together like reasoning chains. Deductive reasoning (from premises to conclusions) can be used for predicting reaction outcomes. Conversely, abductive reasoning (identifying possible premises given the conclusion) can be used for retrosynthetic analysis—identifying possible reactants for synthesizing a given product. Both are high-impact problems in organic chemistry with the potential of generating huge practical value. Compared to neural networks, my reasoning approach has unique advantages on problems in machine learning for scientific discovery. It generates not only predictions but also interpretable symbolic proofs, which are necessary for enhancing our scientific understanding.

I am particularly excited about advancing fundamental AI capabilities while addressing pressing social needs, such as AI for drug discovery, renewable energy, and environmental protection. I look forward to exploring these multidisciplinary problems with collaborators from academia and industry.
References


