Improving Learning-to-Defer Algorithms for Human-AI Collaboration in Autonomous Vehicles

BAA Information

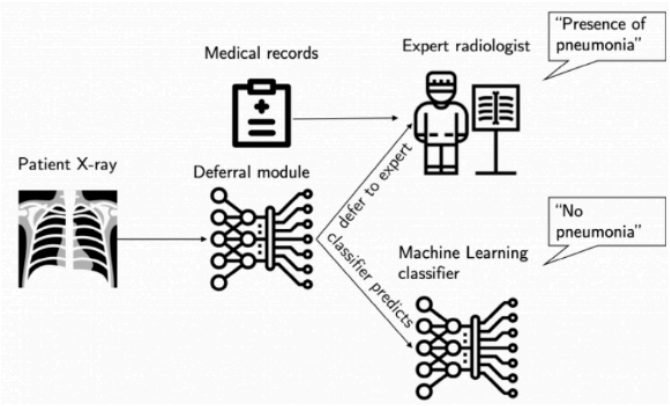
BAA Number: FA9550-21-S-0001 [1]
BAA Title: Air Force Office of Scientific Research Broad Agency Announcement [1]
Relevant ONR Technology Area: Trust and Influence

Key Words: Artificial Intelligence, Human-AI Collaboration, Autonomous Driving

Motivation

Advances in artificial intelligence (AI) have increased automation and improved decision-making, while also using fewer human resources. Despite improvements in AI technology, we still lack a complete understanding of how AI makes decisions due to its black-box nature, leading to problems of bias and unpredictability [2]. AI needs to be understood and trusted before deployment, and so it's important to develop mechanisms that can curb the risks of AI. In this proposal, we focus on the use of humans to curb risks and improve safety for autonomous vehicle systems. Combining humans with AI for autonomous vehicles is difficult due to the need for robust across driving conditions, but previous work has shown that AI can adapt to a wide range of conditions, such as a Department of Defense (DoD) program that uses neural networks to improve sensors for hypersonic vehicles [3]. Developing safer and more trustworthy autonomous vehicles allows for increased automation in the DoD, which can reduce costs.

We can use learning-to-defer algorithms [4] as a way to combine human and AI strengths, reducing AI risks. These algorithms use deep learning and convolutional neural networks to partition tasks between humans and AI depending on which is more suited for a particular task. Learning-to-defer algorithms are dependent on a human model, which asses the difficulty of a particular task for humans, and uses that score, along with AI predictions, to determine whether to defer to humans. We propose improving learning-to-defer algorithms for autonomous vehicles by studying differences in skill between humans and AI. We do this by tackling the following three questions



Learning-to-defer algorithms leverage experts to make predictions (figure from [4])

1. **RQ1:** How can we quantify differences between human and AI driving skills?
2. **RQ2:** How can we incorporate fine-tuning into learning-to-defer algorithms?
3. **RQ3:** Can we improve out-of-domain generalization through upgraded vision models?

Understanding the failure modes of AI and humans and developing learning-to-defer algorithms using this information is useful for the DoD. An understanding of failure modes allows for improved safety through a more accurate assessment of potential failures, leading to better mitigation strategies. Additionally, improved learning-to-defer algorithms can be transferred to

other areas within the DoD, such as control systems, which maintain conditions in various settings such as aircrafts. For control systems, learning-to-defer algorithms can address abnormal situations where human takeover is necessary, curbing the negative impacts of AI decisions, and potentially alerting humans about dangerous situations, such as cybersecurity breaches [5].

RQ1: How can we quantify differences between human and AI driving skills?

To better understand autonomous vehicles, we propose employing an item response theory (IRT) model to quantify the differences between human and AI driving abilities. IRT models quantify skill by expressing it as a function of competency and task difficulty [6], thereby taking into account task difficulty when predicting an agent's skill. These models are prevalent within education to determine test taker skills; while some questions are answered by everyone and others by no one, medium-difficulty questions can distinguish between students. IRT models have also been applied to question answering [7], to distinguish between question answering models based on their performance on different data points. In the context of autonomous vehicles, IRT models can help distinguish skill levels between humans and AI and determine what conditions each excel at. These conditions can include environmental properties, such as weather and time of day, in addition to factors like trip duration.

To run an IRT model, we need information on human and AI driving capabilities, and so we run a simulation where humans and AI drive a simulated car over a closed course. To assess performance, we develop safety metrics, such as braking times, lane departure frequency, and disengagement rate (a measure of how often humans need to intervene for AI). Using this data, we would apply an IRT model to understand the conditions under which it's advantageous for humans to drive over AI. The IRT model fits to simulation data, essentially performing regression, and these regression parameters can inform us of the difficulty of various situations for humans and AI. This can inform learning-to-defer algorithms on when to defer, and generally helps us understand AI driving capabilities better.

RQ2: How can we incorporate fine-tuning into learning-to-defer algorithms?

We can use the results from our IRT study, along with fine-tuning, to improve human models for learning-to-defer algorithms. Learning-to-defer algorithms use human models to assess the difficulty of tasks for humans and determine whether humans or AI should solve a task. Both the learning-to-defer Improved human models, which assess human skills, could result in improved learning-to-defer algorithms. We can improve human models through knowledge of the individuals AI works with. For example, if an autonomous vehicle is working with a driver who is good at driving in the rain but not in the snow, then AI should defer under rainy circumstances, and not defer under snowy ones. In general, fine-tuning accounts for the specific individuals AI works with and can improve learning-to-defer algorithms. To fine-tune human models, we should take advantage of both information about the individual that AI will work with, and general information on driving skills for all humans. This means AI should understand that in general, humans are worse at driving in the rain vs. sun, and that the individual they drive with might have particular strengths and weaknesses on top of that. One strategy to incorporate this information is to train using individual-specific information and treat general information as unlabeled data which can inform the underlying manifold. The manifold gives information on where data points, which represent driving conditions, can reside. Learning from both labeled and unlabeled data can be done by employing semi-supervised learning algorithms.

My prior work at a federally funded research and design center located in Massachusetts showed that self-training, a semi-supervised learning algorithm that trains on labeled data and imputes predictions on unlabeled data, can improve learning-to-defer algorithms through fine-tuning [8]. One way to build on this is to change the objective function by adding an entropy regularization term. Entropy regularization encourages similar, unlabeled points to have similar predictions, taking advantage of unlabeled points in optimization. Additionally, we can incorporate information from the IRT studies to determine which features are most indicative of AI vs. human skill, using this information to perform feature selection. The IRT studies would additionally give information on the underlying data manifold, which indicates the relationship between driving conditions and human performance. To take advantage of this, we could employ graph-based semi-supervised learning, which uses the similarity between data points as an additional objective for optimization. Combining semi-supervised learning with the IRT study can help fine-tune models, resulting in better learning-to-defer algorithms.

RQ3: Can we improve out-of-domain generalization through upgraded vision models?

Improved vision models can be combined with semi-supervised learning so learning-to-defer models generalize to out-of-domain data, such as unseen environmental conditions. Out-of-domain generalization is important, as AI models should make predictable or reasonable decisions when encountering new data points. This is especially important within the DoD, as improved out-of-generalization increases trust in AI.

We can use vision transformers [9] to improve out-of-domain generalization. Vision transformers are based on transformers from natural language processing (NLP) [10], and they use an attention mechanism to weigh the importance of different parts of the input. Prior work has shown that vision transformers improve out-of-domain generalization when compared to traditional convolutional neural networks (CNN) [11]. Learning-to-defer algorithms for autonomous vehicles use CNNs to learn whether to defer from images of environmental conditions. To incorporate vision transformers, we could either replace CNNs directly with vision transformers, or incorporate self-attention mechanisms as an addition to CNNs. We would then need to test out different architectures, loss functions, and hyperparameter combinations to determine an optimal combination that improves decision accuracy. To assess the impact of our vision transformers, we would evaluate learning-to-defer algorithms using simulation data, training on some in-domain data such as sunny days, and testing on some out-of-domain data, such as rainy days. Improved vision models would be widely applicable across the DoD, and can be used for other projects, such as Automated Aerial Refueling (AAR) estimation, which aims to find 3D distances from 2D images.

Summary of Proposal

Our proposal aims to improve learning-to-defer algorithms for autonomous vehicles, so AI can efficiently decide whether humans or AI should drive a vehicle. We do this by studying differences between human and AI driving skills, then use this in conjunction with fine-tuning to make learning-to-defer algorithms individual-specific. To improve out-of-domain generalization, we propose the incorporation of vision transformers into learning-to-defer algorithm architectures. Improving learning-to-defer algorithms increases automation while improving the trustworthiness of AI, allowing for expanded AI use within the DoD. Human-AI collaboration research can be applied to other areas outside of autonomous driving, such as control systems. Pursuing these research questions can allow us to better understand human-AI collaboration.

References

[1] AFOSR BAA Announcement #FA9550-21-S-0001: Broad Agency Announcement (BAA) for Air Force Office of Scientific Research

[2] Garvie, Clare, and Jonathan Frankle. "Facial-recognition software might have a racial bias problem." *The Atlantic* 7 (2016).

[3] Scott, Katie. "AFIT ENGINEER."

[4] Mozannar, Hussein et al. "Consistent estimators for learning to defer to an expert."ICML. PMLR, 2020.

[5] "Cybersecuring DOD Control Systems." *National Initiative for Cybersecurity Careers and Studies*, https://niccs.cisa.gov/training/search/pmc-group-llc/cybersecuring-dod-control-systems.

[6] Embretson, Susan E., and Steven P. Reise. *Item response theory*. Psychology Press, 2013.

[7] Rodriguez, Pedro, et al. "Evaluation Examples Are Not Equally Informative: How Should That Change NLP Leaderboards?." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021.

[8] Raman, Naveen and Yee, Michael. "Improving Learning-to-Defer algorithms through fine-tuning", *Workshop on Human-Machine Decisions at Conference on Neural Information Systems 2021*

[9] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv preprint arXiv:2010.11929* (2020).

[10] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[11] Zhang, Chongzhi, et al. "Delving Deep into the Generalization of Vision Transformers under Distribution Shifts." *arXiv preprint arXiv:2106.07617* (2021).

Personal Statement

Improvements in artificial intelligence (AI) and machine learning (ML) algorithms both excite and scare me. On one hand are applications like the kidney exchange program at Facebook, which uses big data to deliver life-changing results, and on the other are racial bias issues surrounding the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) recidivism algorithm. I aspire to work on AI and ML for social good, and my interest comes from both a technical fascination with these technologies and a civic belief in the need to use AI and ML to improve society. Developing and understanding AI and ML algorithms requires an interdisciplinary approach, and so I've taken classes on education policy, natural resource economics, and ethics.

My prior research focused on improving the utility and equity of intelligent computing systems, including work on ride-pooling fairness, open-source toxicity, and human-AI collaboration. By trying a variety of ML-based projects, I learned that I enjoy public-facing research that combined theory and application. As a computer science PhD student, I plan to continue researching uses for AI and ML, while finding methods to minimize their bias, specifically for applications involving healthcare and public health.

I aspire to study public-facing applications of ML, with an emphasis on social good. Examples of applications include the use of ML for better clinical prognosis across racial groups and investigations into the use of ML in criminal justice applications. I would be excited to pursue these directions at graduate schools such as MIT and Harvard. At MIT, I would love to join the Clinical Machine Learning lab, which researches how ML algorithms can improve clinical prognosis while ensuring the fairness of those algorithms. I would also be excited to join Teamcore at Harvard, which uses AI to design public health interventions. My goal is to produce impactful research during my PhD, and to this end, I plan to work with organizations to assist with their use of ML and ensure its equitable use. Examples include working with World Wildlife Federation to develop ML algorithms that position security guards to minimize poaching or working with Safe Place for Youth to develop interventions that distribute HIV prevention information.

After graduate school, I plan to become a professor, researching applications of AI and ML, and working with local organizations to ensure the equitable use of AI. My experience as a teaching assistant allowed me to discover my enjoyment of teaching, especially with one-on-one and small-group discussions, so I aim to continue teaching courses and practicing my communication skills in graduate school. Becoming a professor allows me to combine my interests in teaching and research.

Receiving the NDSEG fellowship would allow me to meet talented young scientists, with whom I could collaborate on interdisciplinary projects. The fellowship additionally provides funding for my PhD, allowing me to focus on research rather than my source of funding. Finally, I get the opportunity to travel, both through the travel stipend for conferences and to the NDSEG conference where I can meet other fellowship winners.