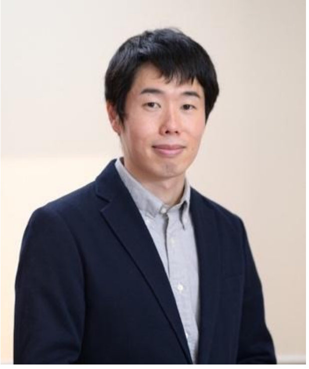
ATML Fellow Class 2025

Dr. Takashi Ishida

RIKEN AIP and University of Tokyo

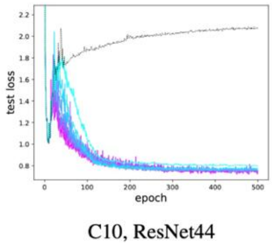
https://takashiishida.github.io



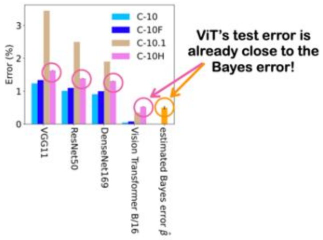
Bio: Dr. Takashi Ishida is a Research Scientist in RIKEN AIP and a Lecturer in The University of Tokyo. At UTokyo, he is co-running Machine Learning and Statistical Data Analysis Lab. He earned his PhD from the University of Tokyo in 2021, advised by Prof. Masashi Sugiyama. During his PhD, he completed an Applied Scientist Internship at Amazon.com (hosted by Dr. Amjad Abu-Jbara) and was a PhD Fellow at Google (mentored by Dr. David Ha) & Research Fellow at JSPS (DC2). He is interested in data evaluation, weakly supervised learning, and distribution shift. He served as an Area Chair for ICML and ICLR, an action editor for TMLR, and reviewer/PC for NeurIPS, ICML, ICLR, AISTATS, and others. He received the Funai Information Technology Award for Young Researchers in 2023.

Contribution:

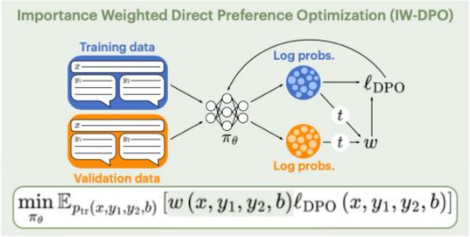
Building robust machine learning systems requires addressing incomplete supervision, ensuring generalization under limited data or distribution shifts, and measuring theoretical performance bounds. Our research starts with weakly supervised learning approaches: we introduced positive-confidence learning, which trains a binary classifier solely on positive data attached with a confidence score (without needing any negative or unlabeled data), and learning from complementary labels, a multi-class approach where each label indicates only which class an instance does not belong to. Both methods reduce the need of fully labeled datasets by leveraging forms of supervision that are easier to obtain. To combat overfitting when data are insufficient, we proposed flooding, which deliberately stabilizes the training loss at a modest level to avoid overconfidence. Once these issues are addressed, we can approach the theoretical limits of classification performance, motivating the creation of a model-free, instance-free Bayes error estimator that helps determine how close we are to the best possible accuracy and whether further tuning or data collection is worthwhile. Finally, real-world deployments often involve distribution shifts that can significantly harm alignment strategies such as reinforcement learning from human feedback, and we propose a robust method that can keep models aligned with user prompts and preferences.


C10, ResNet44

Better performance when flooding is used.



Comparing the estimated Bayes error with model performance.



$$\min_{\pi_\theta} \mathbb{E}_{p_{\text{tr}}(x,y_1,y_2,b)} \left[w\left(x, y_1, y_2, b\right) \ell_{\text{DPO}}\left(x, y_1, y_2, b\right) \right]$$

Addressing distribution shift in language model alignment.

Dr. Thai Le

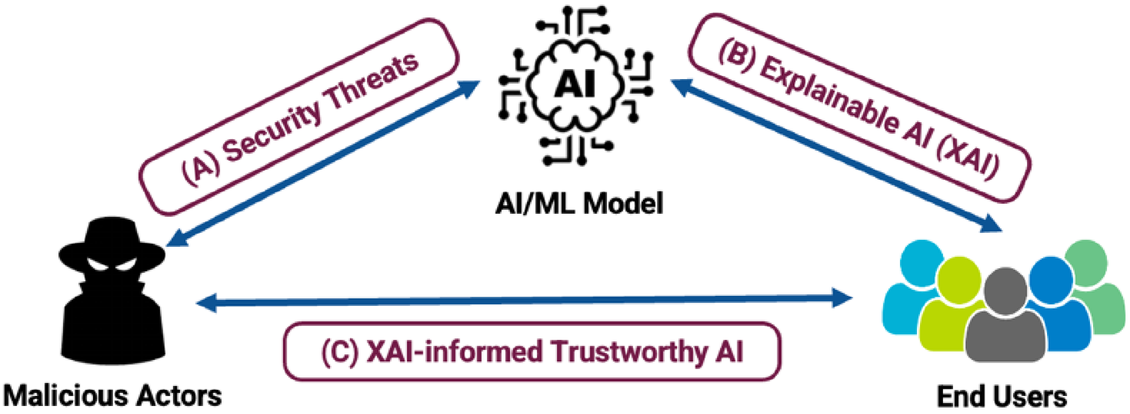
Indiana University

https://lethaiq.github.io/

Bio: Dr. Thai Le is an Assistant Professor of Computer Science at the Indiana University Luddy School of Informatics, Computing, and Engineering. Before joining Indiana University, he was an Assistant Professor at the University of Mississippi and had industry experience at Yahoo Research and Amazon Alexa. He earned his Ph.D. from the College of Information Sciences and Technology at Penn State University under the supervision of Prof. Dongwon Lee, receiving a Doctorate Excellent Research Award and a DAAD AInet Fellowship. Dr. Le's research focuses on the trustworthiness of AI/ML models, particularly in Natural Language Processing (NLP). His mission is to enhance the robustness, safety, and transparency of AI technologies in various sociotechnical contexts, ensuring that society and internet users can harness their power with confidence and clarity. Dr. Le has published over 40 peer-reviewed research papers in top-tier AI/ML and NLP venues, including AAMAS, AAAI, ACL, EMNLP, NAACL, KDD, and CHI. His research has been featured in ScienceDaily, Defense One, and Engineering and Technology Magazine.

Contribution:

Dr. Le has conducted extensive research on security for AI (A), particularly in the context of national security. He has pioneered several novel algorithms, including MALCOM, which generates convincing yet malicious comments on social media posts to deceive state-of-the-art deceptive information detection systems at that time with over 90%—even predating the rise of transformer-based models. He also developed ACORN, which leverages reinforcement learning to control social bots to spread fake news while evading state-of-the-art detection methods, raising awareness of how malicious actors can utilize AI to amplify online deception. He invented DARCY, the first algorithmic honeypot for NLP designed to trap potential universal triggers by artificially injecting local extrema on the loss landscapes of textual neural network models, enabling a more proactive defense approach against adversarial text attacks. Additionally, Dr. Le pioneered the first and largest dataset of real-world text perturbations, called ANTHRO, enabling the study of adversarial robustness in NLP models against inductive, human-written (rather than deductive, machine-generated) perturbations. Researchers across nearly 30 countries have since utilized this dataset to understand better adversarial behaviors in online texts and benchmark state-of-the-art generative models such as Dall-E with natural, noisy texts. Dr. Le and his collaborators were among the early researchers to explore deepfake text detection, beginning in 2017—a field that has become increasingly relevant with the rise of ChatGPT and other generative AI technologies.



Dr. Le has also contributed to explainable AI research (B), including GRACE, an early algorithm on counterfactual explanation, and NoMatterXAI. This novel algorithm generates alterfactual explanations to explain NLP classification models and help illustrate the bias of these models to the end-users. Together with his interdisciplinary collaborators, Dr. Le also extended these algorithms to explain complex reinforcement learning algorithms and their safety mechanisms in various contexts. Ultimately, these works will help us defend and protect AI models against future threats (C).