Stanford University

November 15, 2024

To whom it may concern,

I am very pleased to write this letter in support of Min Wu. Min is a postdoctoral scholar at Stanford University, and I am her research advisor. Min works in the exciting and important area of *AI Safety*. In particular, she works on applying formal methods to machine learning systems to improve their reliability and explainability.

Min came to Stanford in early 2022, following a very successful PhD at Oxford with Prof. Marta Kwiatkowska. There, she participated in a number of groundbreaking projects applying formal methods in machine learning. Since I was not part of that work, I will leave it to Min or Prof. Kwiatkowska to describe that work.

Since coming to Stanford, Min has continued to contribute to this research direction, while also taking an important leadership role in the Stanford Center for AI Safety, both of which I will describe below.

Before explaining further, let me first give some brief background on formal methods applied to machine learning. My lab's work in this direction began in 2016, when Professor Mykel Kochenderfer, a colleague of mine in aeronautical engineering, was working with the United States Federal Aviation Administration (FAA) to develop an automated collision avoidance system for unmanned aircraft called ACAS Xu. For various performance reasons, they were considering using a trained neural network as a controller in this system. However, not surprisingly, people were nervous about putting a neural network in a safety-critical system. Mykel asked me if we could formally verify the controller. This seemed like a timely and interesting challenge, so I agreed to try. We quickly discovered that existing solvers—both Satisfiability Modulo Theories (SMT) solvers and linear programming (LP) solvers—were not up to the task, so we started talking about designing a new custom solver. The result was the Reluplex solver, which was able to prove many properties of the ACAS Xu networks. This work, presented at CAV 2017, effectively kicked off a surge of research in *neural network verification*, which continues to this day. As one confirmation of the importance of this new research direction, my co-authors and I recently received the 2024 CAV award for our work on Reluplex.

Min comes from another lab that helped pioneer this research direction. She arrived at Stanford with energy and enthusiasm and proceeded to learn about and get involved in all of the work happening here. One of the first projects Min proposed to me was to investigate formal approaches to explainability. Several approaches for explainability already existed in the literature, but they were either based on heuristics (providing no guarantees) or were limited in scope (targetting only NLP applications, for example). Min proposed applying our formal tools to provide formal guarantees about explanations for perception networks. The idea is as follows. Given a machine learning system for, say, classification, for a given input, provide a subset of the features in that input that are sufficient to *guarantee* that the system will classify the input in the same way. Min developed the theory and code for this approach, calling it VeriX

(for verified explanations). This work appeared in NeurIPS 2023. A second paper which extends and optimizes this work, and also demonstrates how it can be applied to out-of-distribution detection, is currently under review at AAAI.

Min has also taken the lead in a collaborative project with IBM, looking at applying formal methods to foundation models. We started by tackling the problem of formally reasoning about softmax and bilinear network layers. Min was a crucial contributor to a paper providing new approximation bounds for softmax that was presented at AISTATS 2023. A paper on approximation bounds for the bilinear function is currently in preparation.

Min has also contributed to several other efforts in the lab, including a new approach to AI planning based on using a similar approach to that used in neural network verification (published at IROS 2023), an effort to verify properties of *quantized* neural networks (publications at AAAI 2024 and SAIV 2024), and a new release of the Marabou neural network verification tool (paper published at CAV 2024). She continues to be involved in all of these projects, and I expect many more excellent results are forthcoming.

When I hired Min, I asked her to take a leadership role in the Stanford Center for AI Safety. The center is a hub of activity that brings together a variety of faculty from different departments. I was hoping to foster more communication and a better sense of community in the center. Min happily accepted this additional role and responsibility. She led weekly meetings on community building that brought together students and postdocs from various labs. She planned a series of events, including lunches, lightning research talks, and seminars. She also helped with the annual meeting bringing together researchers and industry affiliates. Finally, she has played a key role in writing funding proposals for the center. In short, Min has become an essential part of keeping the center functioning and thriving.

I want to emphasize a few points related to the work that Min has done. First of all, Min has a deep expertise in both machine learning and formal methods. This is quite unusual, and makes her extremely valuable as a collaborator on projects. She is able to just as easily write about formal theory as run experiments to train machine learning models. Second, Min is hard-working and reliable. Whenever she is given a task, I have complete confidence that it will be finished professionally and on time. She has been invaluable for developing presentations for the center that are polished and clear. Third, Min works really well as a team member. She contributes selflessly and communicates well with collaborators.

Finally, Min has recently taken on an additional role as an undergraduate research mentor. She proposed and defined a project, interviewed prospective applicants, and helped select the student we ultimately hired. She has since helped that student develop the necessary skills to succeed, recommending courses and assisting them with their course projects. The student successfully contributed to research and is a co-author on our AAAI submission.

To summarize, Min is an outstanding researcher embedded and involved in all aspects of the rich research environment at Stanford. She has worked hard to develop all of the skills that are necessary for a successful academic career. Her accomplishments are similar to those of other top researchers in AI Safety at a similar stage, e.g., Guy Katz (professor at Hebrew University), Gagandeep Singh (professor at UIUC), and Zico Kolter (professor at CMU). You cannot go wrong

by betting on Min. She is the full package: a machine learning expert and a formal methods expert; she is smart, creative, ambitious, and productive. She has broad interests and is well on her way to becoming a superstar. I expect she will help define future research areas that merge formal methods and machine learning. I feel privileged to get to work with her as she makes this journey. She has my highest recommendation.

Sincerely,

Clark Barrett, *Professor (Research)*

Department of Computer Science
353 Jane Stanford Way, Stanford, CA 94305 • barrett@cs.stanford.edu • http://cs.stanford.edu/~barrett