Caleb Schultz Kisby

Email: cckisby@gmail.com **Website:** ais-climber.github.io

Position: Postdoctoral Research Position in Human(e) AI, University of Amsterdam

Dear members of the Human(e) AI Steering Board,

I'm writing to apply for a Postdoc position through your Human(e) AI RPA, within the AI and Logic track. I am currently a PhD candidate in Computer Science at Indiana University (expected defense: April 2025). My research focuses on foundational questions that underlie artificial intelligence (AI) and cognition, in particular:

- How should we best integrate symbolic and neural (sub-symbolic) systems?
- How can we extract, interpret, and verify the internal beliefs of neural networks?
- How powerful and reliable are different learning algorithms, when compared to one another?
- Is provably correct AI alignment possible?

I tend to approach these questions using tools from logic and theoretical computer science. But this work is necessarily interdisciplinary in nature, so I also borrow and share ideas across many other fields including machine learning, philosophy, psychology, linguistics, and neuroscience. As my CV shows, I contribute to both mainstream AI conferences as well as to interdisciplinary meetings (including the cognitive lunch seminar at IU, the LIRa seminar at UvA). For my PhD, I have answered many of these questions in a somewhat simplified setting. My long-term goal is to see these questions answered for neural networks and learning algorithms that are used in practice.

I have in mind two points of collaboration within your Human(e) AI initiative. First, I'm interested in working with the Amsterdam Dynamics Group at the ILLC on developing formal logics with the aim of designing safe, trustworthy, and interpretable machine learning systems. Actually, Sonja Smets encouraged me to apply to this position in the first place—we met in January, along with Alexandru Baltag, and realized that my approach to modelling dynamics in neural networks bears striking resemblance to her work in modelling dynamics in *social* networks. On a personal note, conversations with Sonja and Alexandru, as well as the work of the Dynamics group as a whole, have had a big influence on my development as a researcher. I would be delighted to have the opportunity to work with this team.

Outside of AI and logic, I'm interested in learning from and sharing ideas with with the Epistemological and Ethical 'Explainable AI' team. As I explain in my proposal, I plan on implementing agents that obey constraints *before and after* they learn from data. A crucial step here is to come up with realistic, human-interpretable ethical constraints and formalize them in the logic language. As far as I'm concerned this requires the expertise of a team like yours, and I would be very happy to be in a position to work together on this.

Thank you for your time and consideration. If you have any further questions, I'm available at the email above, as well as over Zoom.

Caleb Schultz Kisby