Expression of Interest for CSCW 2025 Workshop on Responsibly Training Foundation Models

Luis Galárraga
Inria, France

I am Luis Galárraga, full time researcher at Inria (Rennes, France) working on knowledge representation and on the algorithmic and human aspects of explainable AI (XAI). My ultimate research goal is to make AI systems fully self-explainable, which implies to make the logic of AI systems fully understandable to humans, but also to guarantee that such explanations reflect what is actually going on inside an AI system.

For a researcher working on eXplainable AI (XAI), all the steps and aspects in the deployment of AI models are crucial, because they have a direct impact on the (a) goal of applying explainable AI techniques, and (b) the expected properties of the explanations. This is even more crucial for complex foundation models that are usually full-blown black boxes and that constitute an important brick within large and complex AI systems. In that light of thought I am convinced that this workshop will help me obtain a more informed view of all the ethical and technical considerations in the construction of training datasets for foundation models – a field in which I am still not very knowledgeable but I am willing to discover. Conversely, I believe that my experience in machine learning, explainable AI, and human-centered AI will be an asset in the discussions as it will provide the perspective of the ML practitioner who is constantly making technical and architectural decisions.