Been Kim, AI Interpretability. Staff Research Engineer at Google Brain, March 2023

Been Kim, expressed her desire to ensure that AI benefits humans, and that we should act towards this goal now rather than later. She emphasized the importance of accountability and proposed treating AI like a colleague, encouraging better communication between machines and humans.

Kim presented a framework of three questions to guide the investigation of AI interpretability: 1) have interpretability tools been built on incorrect assumptions? 2) Are our expectations of what they do incorrect? and 3) Are AI systems beyond our understanding? For each question, she offered an approach to investigate it, such as studying machines through observational and controlled studies and studying how humans use AI.

Kim also scrutinized existing interpretability tools and concluded that the Shapley metric and saliency maps are no better than random chance in some cases. She also investigated the possibility of discovering emergent behavior in multi-agent systems and highlighted the need for testing such systems to address concerns of potential harm. Finally, she discussed her recent work on using superhuman AI systems to retrain humans.

Overall, Kim's message is that AI interpretability is essential for developing the next generation of methods to test AI systems and ensure their accountability and safety.