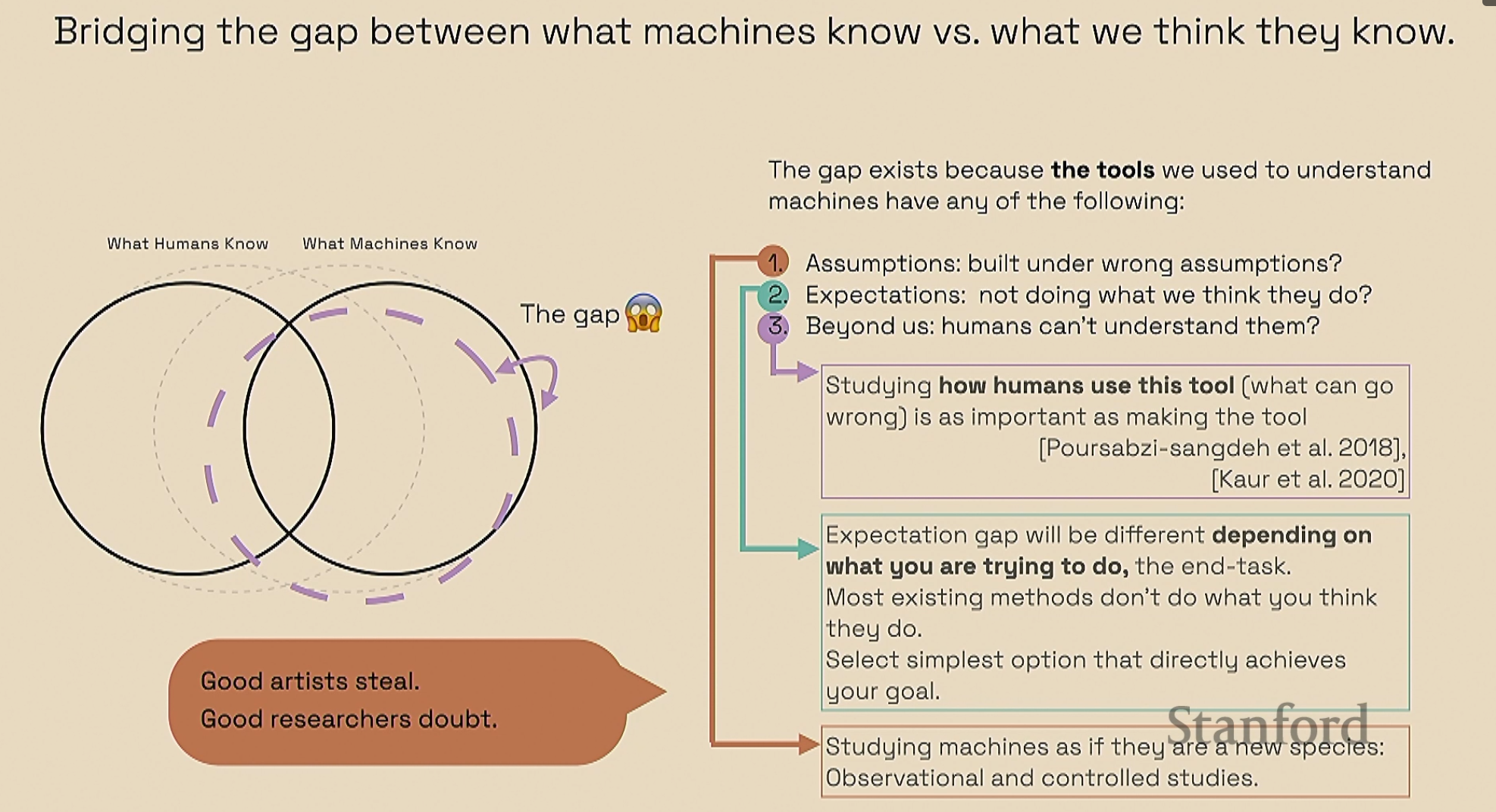
**Been Kim, AI Interpretability. Staff Research Engineer at Google Brain, March 2023**

Been affirms what we are all thinking: *“Not sure where we’re going with this technology”.* She said her *“hopes and dreams”* are to have AI that benefits humans and to do it now – why wait? And said that her young son will keep her accountable. This was profound for me. I’m expecting my own kid this week. Sounds corny but I need to choose how I spend my time and career to make the world better for her.

Been proposed to treat AI as a colleague by trying to have a conversation and have the AI explain itself to us. *“Solving this problem is a superset of solving AI safety”,* which implies it’s harder. She wants to build machines that can better communicate with us. What would that look like? Is it like asking ChatGPT to explain how it got its answers? How can I trust that answer?

She has an assessment framework comprises three question: 1) did we build them on the wrong assumptions? 2) are our expectations about what they do is wrong? 3) maybe it’s superhuman and beyond our understanding?. For each, she has an experimental approach 1) study machines as if they are a new species via observational and controlled studies. This is akin to mechanical engineering, where the machines informed the theory, 2) select the simplest tool to achieve your goal, 3) study how humans use the AI.

Been scrutinized existing interpretability tools and concluded that shapley metric and saliency maps are no better than random chance. She investigated how to discover emergent behaviour in multi-agent systems? The goal is to design intervention systems that perform as well as the baseline. This was aid safety assurance of cobotic swarms in unstructured environments like the home. Finally she concluded with her work on feeding back the skills from superhuman AI systems to humans. That’s very cool.



She showed that, counter to practice, the theory shows that saliency maps of trained vs. untrained models are not better than random chance. She concluded the same, unless I’m mistaken, about Shapley metric. She claims that we can’t “infer model behaviour with popular feature attribution methods in theory and practice. She doesn’t dismiss it altogether and acknowledges the tools are imperfect but with human in the loop it’s possible to be ok.

She strong advocates simple methods over complex. I really want to take this to heart.

Further readings

1. Further reading: Beyond interpretability: developing a language to shape our relationships with AI ICLR2022 Keynote
2. Saliency map is used for computer vision. Random prediction

Keywords

1. Saliency Map in Computer Vision
2. Shapley metric
3. Gestalt Phenomenon 2021