Popular science summary of the PhD thesis

PhD student	Beatrix Miranda Ginn Nielsen
Title of the PhD thesis	On Representations of Generative Models
PhD school/Department	DTU Compute

Science summary

Imagine someone gives you a black box which they say can answer every question you might have. Just to test the box you ask it: "What is one plus one?" and it replies "two". Great! You think but let us just try a slightly more difficult question. This time, you ask: "What is one plus one? Oh, and did you know that cats sleep for about 20 hours a day?" To this the box answers: "Three". This might seem weird to you, so you might ask the people who made the box why it gave a wrong answer the second time. How would you feel about using the box, if the makers of the box told you "We don't know"?

With black boxes such as ChatGPT, AI models have become mainstream. These models give very natural sounding replies to any question you might ask them, but they also get confused by facts about cats and say something wrong, and even the people who make them cannot properly explain how a model came up with an answer to your specific question.

These models cannot read sentences, so what they do is convert the sentences into a bunch of numbers. They then use the numbers to do some calculations, and they take the results of these calculations and convert them back into text so that humans can read their reply. This bunch of numbers which the models use to calculate their answers we call their *representations*, because they use the numbers to *represent* sentences since they cannot do calculations on text.

I think that if we are ever going to understand how these models work, we need to understand the representations the models use, because these representations are the only way the models ever "see" and interact with the world.

My PhD is about understanding the representations of AI models. So, I ask things like: How does a model compare its representations to each other when doing calculations? And which representations does it make sense for *us* to compare if we want learn something about the model? Can we make representations better for the kind of calculations the model does? And can we say something about the representations of two models by comparing their outputs?

In time, I hope work like mine will let us understand how AI models work, so that we can trust them, or at least know when and for what we can trust them.

Please submit the summary to the department PhD coordinator together with your thesis