Excerpt from the Leonard Tang interview

Leonard: The bill is both simultaneously extremely specific about what constitutes harmful and not harmful models, or what is in and out of scope of the bill, yet is extremely under-specified about the sort of things it actually is trying to test for. In the bill, it does directly ask for third-party testing for safety categories, but they never really specify what categories we should be testing for or to what quantity and magnitude we should be getting in terms of coverage or in terms of the quantity of prompts that we're sending to the model. These things make a meaningful, extremely meaningful difference in terms of the behavior of your model. If you're a couple dozen prompts off, that could mean the difference between eliciting a particular behavior or not eliciting a particular behavior. I think it's extremely fuzzy to try and quantify what is reasonable coverage and not coverage. The downstream application of this, then, is I think it'll not be so much very logical in the way that people approach safety testing. It'll be more just whoever has the most influence and sway and loudest voice in the room gets to call the shots on what is reasonable and not reasonable.

Michaël: One thing you said is that testing, red teaming, and mitigations are best done with respect to the end users' context and domain. What do you mean by "with respect to the end users"?

Leonard: I think it's sort of unreasonable and almost impossible to try and regulate a priori at the base model layer what the user should be protected against. For example, there are very legitimate use cases for certain types of behaviors or models. Imagine you're an AI companion or an AI artist or creator tool or something like this. It is extremely subtle and nuanced what is considered, let's say, NSFW there and what is considered safe and not safe vis-à-vis, let's say, an AI for financial enterprise or something like this. In the latter category, there are just extremely explicit guidelines for what is acceptable or not. But if you're an AI companion, you really need to be much more nuanced about what is and isn't acceptable in terms of content and behavior.