

THE RESEARCH UNIVERSITY PRESS

International
Computational Law Journal

VOLUME 2 / ISSUE 2 / CURATED SOURCE EDITION

Regulate Systems, Not Models

Kevin Riggle

MIT Computational Law Report · Originally published 2024-09-10

Independent reader edition under CC BY 4.0. This attribution cover was added; original source PDF pages follow unchanged. Original authorship and source identity are retained. MIT, Stanford University, Stanford Law School and CodeX have not sponsored or endorsed this republication.

Source, license and version evidence accompany the online article.

MIT Computational Law Report

Regulate Systems, Not Models

Kevin Riggle

Published on: Sep 10, 2024

URL: https://law.mit.edu/pub/regulatesystemsnotmodels

License: Creative Commons Attribution 4.0 International License (CC-BY 4.0)

Hot on the heels of all this writing I've been doing about AI safety (previously: why the current approach to AI safety is doomed to failure, what AI safety should be, a worked example)— my friend and South Park Commons colleague Derek Slater passed on this California state senate bill, proposed by my senator Scott Weiner, California Senate bill SB1047.

Since Sen. Weiner is my state senator, I've already got a message out to his office to discuss the bill, but having now read the full text in some detail, I wanted to set down some thoughts here for other people as well.

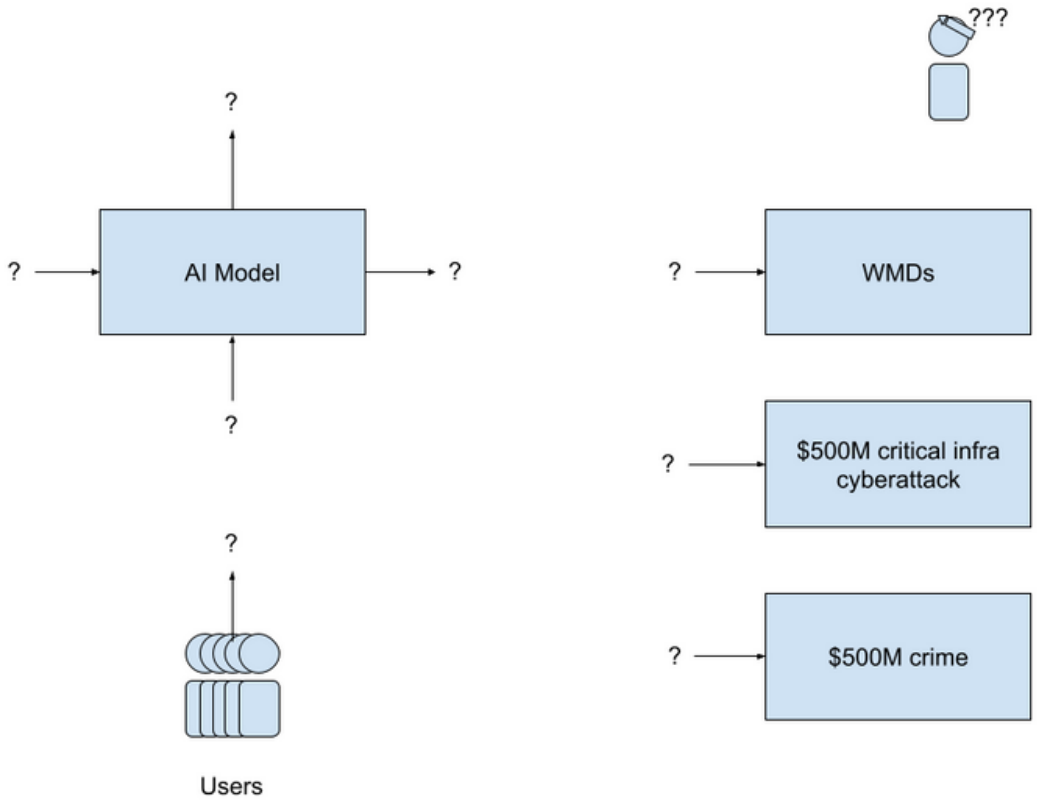
The bill has the mouthful of a title, the "Safe and Secure Innovation for Frontier Artificial Intelligence Systems Act," and, as I read it, I think it breaks down fairly cleanly into two big parts:

- Create a Frontier Model Division within the California state Department of Technology, and require developers who are or intend to train "covered models" to make various certifications about its safety and capability to the Frontier Model Division, with various penalties if they don't, the certifications are later found to be false, or an incident occurs.
  - A covered model is defined as a model trained using a quantity of computing power greater than 10^26 integer or floating-point operations in 2024, or a model that could reasonably be expected to have similar performance or "general capability" to such a model).
  - The certifications are particularly concerned with some abstract "hazardous capabilities" which models might possess, which might make substantially easier than otherwise the specific harms of:
    - "The creation or use of a chemical, biological, radiological, or nuclear weapon in a manner that results in mass casualties."
    - "At least five hundred million dollars ($500,000,000) of damage through cyberattacks on critical infrastructure via a single incident or multiple related incidents."
    - "At least five hundred million dollars ($500,000,000) of damage by an artificial intelligence model that autonomously engages in conduct that would violate the Penal Code if undertaken by a human."
- Instruct the Department of Technology to commission consultants to create and operate a public cloud platform, "to be known as CalCompute, with the primary focus of conducting research into the safe and secure deployment of large-scale artificial intelligence models and fostering equitable innovation."

I'm going to leave aside the CalCompute part of the bill for the moment, because despite the ostensible focus on safe AI development, my read is that it's just much more an implementation detail, and one that honestly feels kind of tacked-on to me. Let's focus here primarily on the Frontier Model Division and the requirements around it.

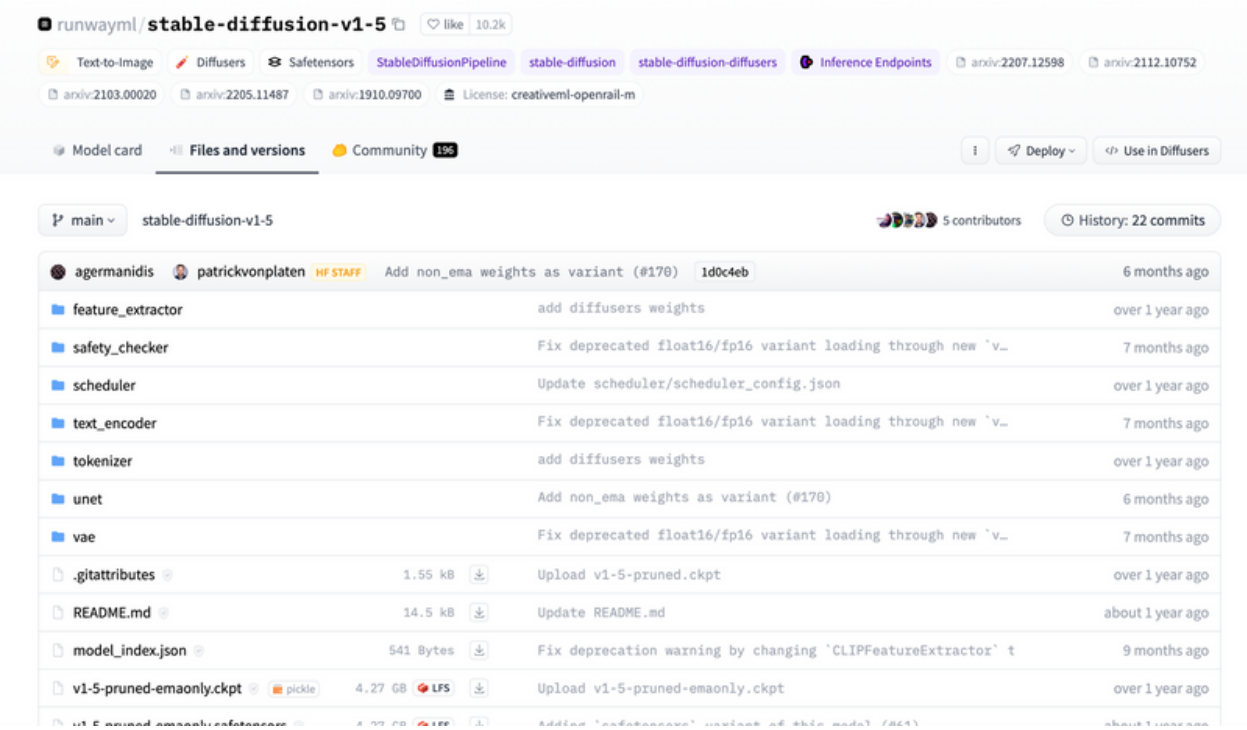
First of all, I like that we're at least defining our unacceptable losses here. I think much about how some fact about the model eventually results in one of these losses is unclear to the point of fuzzy-headedness and too much science fiction, but at least we're putting some kind of stake in the ground about what our cares and

concerns are. As I've written before, this is necessary for us to actually reason about the safety of systems making use of AI models.



A confused system designer surveys the problem space.

The issue here, as I've now written about at some length, is the same as the issue with AI safety writ large.

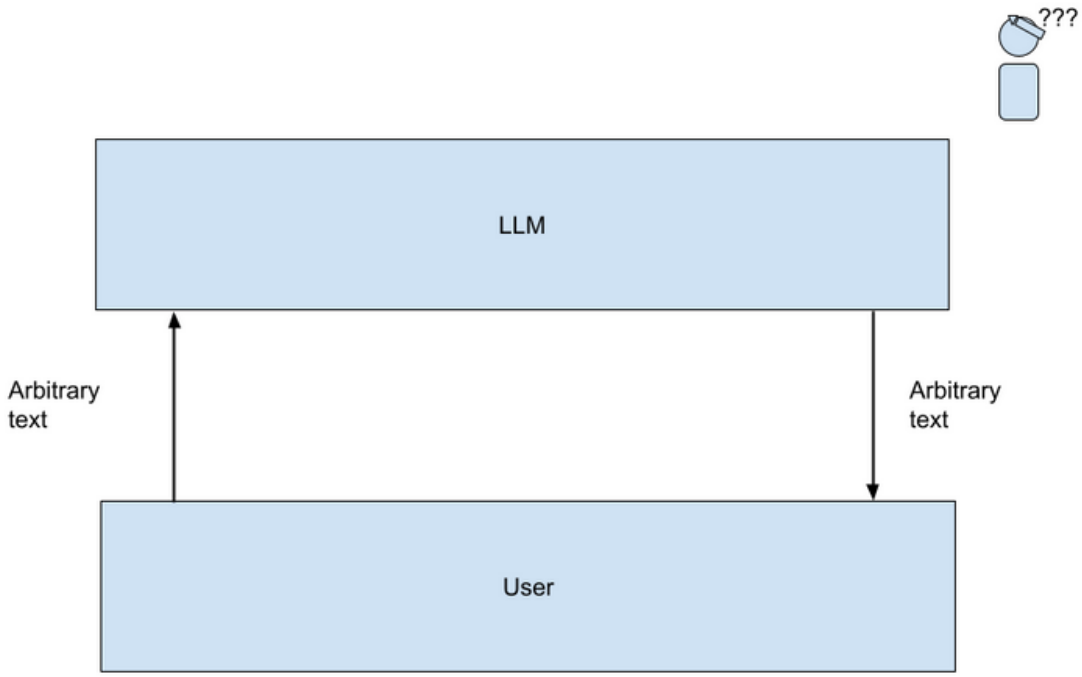


Some of the files comprising the RunwayML distribution of Stable Diffusion v1.5.

AI models don't have capabilities. An AI model is just a file or some files on disk. Systems, whether incorporating AI models or not, have capabilities.

An AI model can't do anything on its own—an AI model can only do things when loaded into a computer's memory, in the context of a computer program or a system of computer programs which connect it to what we would in a control systems context call sensors and actuators, which allow it to do things. (Receive text or images from a user, display text or images to a user, make API calls, etc.)

Those actuators may be connected to other systems (or people), which may be connected to still other systems (or people), which may eventually allow the AI model to be reductively said to have done a thing, but those sensors and actuators are *essential*. Without them the model sits inert.



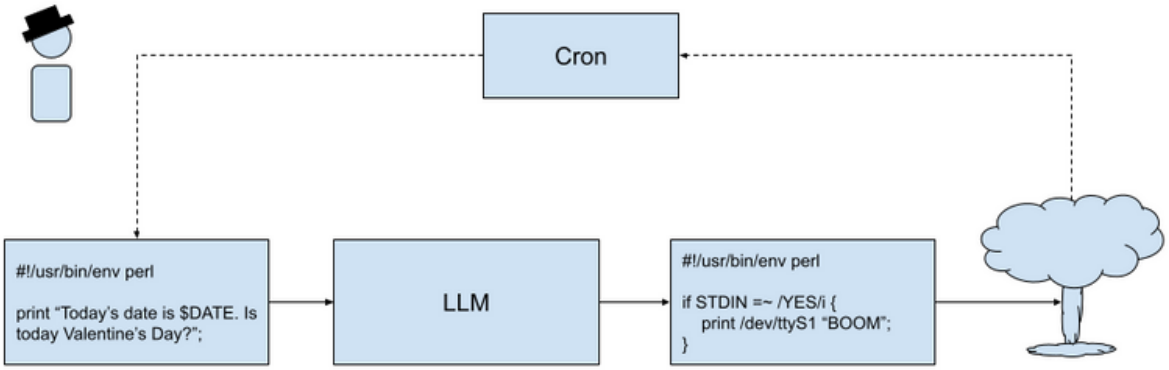
That oversimplified system diagram of a prompted text-generation LLM chatbot, once again.

While it's true that AI developers will talk about their models as having certain kinds of capabilities (e.g. prompted text generation, prompted image generation, image-to-image generation) and as being better or worse at those capabilities on some benchmark than other models, these are only capabilities *within the assumed context of some system*, e.g. a fairly simple system which feeds them text prompts or image prompts and evaluates their output.

I forgot to work my joke about paperclip-maximizer AIs into the first diagram above, and am not attached enough to it to go back and include it, but this gets back to the idea I've mentioned here before that if we don't want an AI model to turn us all into paperclips, one really easy step to take is to not put it in charge of a paperclip factory. (And of course when I go look it up I find that the LessWrong people have decided to rename the concept with a more abstract and less emotionally salient name, haven't they.)

This may all seem a bit pedantic or legalistic, but we are talking about this in the context of a proposed law. Anyway my proposed change to the bill is very simple: We should reword it to talk about constraints on and the compliance of *systems incorporating frontier AI models*, rather than the models themselves.

Now hold on a second, you might say. Doesn't this bring under the jurisdiction of the Frontier Model Division many, many more companies than just frontier model providers like OpenAI and Anthropic? Potentially including anyone deploying their own ChatGPT-powered chatbot? And yes, it does.



An evil system designer builds an evil LLM-containing system. (Please forgive all the syntax errors in the code, these text boxes are small and it's been a long time since I wrote much Perl. You get the point, I hope.)

The only way that OpenAI or Anthropic could even *begin* to consider making the assertions that the Frontier Model Division wants them to make is if they understand and control *very specifically* what systems the users of their models incorporate them into.

One can imagine the above ludicrous but illustrative scenario, where an evil system designer has built a system which asks an LLM "is it Valentine's Day," and, if the system says it is, sets off a nuclear bomb. (Surely a violation of the California Penal Code worth $500M *at least*.) Clearly in this case the AI model *can* in some sense be said to have set off the bomb. And certainly it would be reported in the news as such.

There's just no way for the frontier model provider to know what the downstream effect of an otherwise-innocuous query is in such a scenario or to make the required assertions about it. The people who *can* knowledgeably make the required assertions about the effects the AI models' output could have are the people integrating those AI models into specific systems.

This does not put the foundation model providers out of scope—all of them operate systems like the chatbot system above which connect their AI model to users and potentially a variety of other sensors and actuators. And perhaps scenarios such as my Valentine's Day bomb scenario fall outside the standard of reasonableness.

But by bringing organizations deploying AI models into scope, and regulating them at the system level rather than the model level, we refocus the legislation on the problem we're actually trying to solve and the people best equipped to solve it.

Now the company integrating LLM generation of genetic sequences into its drug-synthesis pipeline needs to assert that they can't accidentally generate a bioweapon, not just the LLM provider. Now the company integrating AI models into their cybersecurity defense platform needs to assert that they can't accidentally DDoS critical infrastructure, not just the company from whom they bought the trained model.

Of course the response to such regulation might be for the frontier model providers to indemnify their customers and take on this responsibility themselves, as some have already discussed doing for copyright infringement. Such a choice would almost certainly, eventually lead to strong contractual and technical controls about to what uses the models could be put and how they could be integrated into larger systems.

This still puts the focus where it most needs to be, and in fact where it must necessarily be—at the point of use.

I'll say it again: AI models and methods cannot be made safe in the abstract. They can only be made safer in the context of particular systems. The more we understand and leverage this fact, the safer our AI-incorporating systems will be.