Feedback on the Draft Report of the Joint California Policy Working Group on AI Frontier Models

Stuart Russell
Professor of Computer Science, UC Berkeley

Overall this is a very useful discussion and makes a number of good points about evidence and policy. The recommendations are all quite reasonable, but much more is needed.

Global comments

The report could go further in terms of ex ante regulation—in particular, towards requiring safety cases providing high-confidence, justified claims of safety. Some references:

- Statement from IDAIS Beijing, March 2024. https://idais.ai/dialogue/idais-beijing/
- Stuart Russell, Edson Prestes, Mohan Kankanhalli, Jibu Elias, Constanza Gómez Mont, Vilas Dhar, Adrian Weller, Pascale Fung, and Karim Beguir, AI red lines: The opportunities and challenges of setting limits. Emerging Technologies, *World Economic Forum*, 11 March 2025.

Transparency is fine but it does not necessarily modify the technology trajectory, which seems to be leading towards extreme risks. Requiring safety cases means that AI companies must develop technology that can support claims of safety. If they can do this for their current technologies, great; if they cannot, the obvious conclusion is that they have chosen the wrong technology path. As the report's Internet example shows, failure to regulate may well allow this path to become locked in.

Despite usefully debunking some common canards around regulation, the report accepts others uncritically. For example, it frequently talks about the need to "balance risks and benefits." This is a stock phrase, often trotted out by ChatGPT, that conveys a misunderstanding of the underlying relationship. It's a false narrative, suggesting that by allowing more risks (and harms) we can get more benefits. (We also need to consider who gets the benefits and who experiences the harms, but I'll leave others to comment on that important issue!) Case in point: deregulation allowing Boeing to sidestep airworthiness certification for the 737 MAX 8. Outcome: 347 dead, $80 billion loss, ceding US leadership in commercial aviation to Europe.

A more accurate narrative is that benefits can be realized only when risks are minimized; i.e., minimize risks to allow benefits, rather than maximize risks to allow benefits. Food, transportation, buildings, electricity provide massive benefits because they are safe, and they are safe in large part due to regulation and liability. If they were not safe, the benefits would be far lower. Thus, we need to ensure safety in order to allow the benefits to be realized.

The report also seems to repeat the misconception, often touted by industry: that we cannot regulate AI because it's changing too fast. This is false. The FDA regulates drugs that change all the time using a formula that doesn't change at all: drugs must be safe and effective. Similarly, nuclear regulators require a mean time to failure of X million years *regardless of the technology* *inside the reactor*, because that's the level of safety we need. We would never accept the argument that reaching that level of safety is too hard, so companies should be able to build unsafe reactors in major cities. One just needs to stop thinking about this from the industry point of view; instead, think about it from the point of view of the responsibility to protect human beings. For AI, we can similarly devise requirements for safety that should apply regardless of what the technology looks like.

The document barely mentions liability; it talks only about reducing potential liability via disclosures. It fails to mention that software and software service vendors typically disclaim all liability. For example, Microsoft limits liability to $5 regardless of the level of harm. In many other areas of the economy, liability serves to ensure that products are designed well. The liability framework is simply non-functioning in the case of software. The suit by Delta Airlines against CrowdStrike is an important test case in which CrowdStrike insists that its contract terms disclaim liability for the harm caused to Delta.

Local comments

p2
"agriculture, biology, education, finance, medicine and public health, and transportation" - none of these require general-purpose AI

"technological design and governance choices of policymakers" – In what circumstances should policy makers be making technical design choices? Why not just impose safety requirements?

"4. Policymakers can align incentives ..." this seems to be recommending a particular approach; and the text does not mention incentives.

"Case studies from consumer products and the energy industry reveal the upside ..." - yes, but those industries have liability for harm (cf. Consumer Product Safety Act); software does not. Surely this is an important lesson?

p5
ensure these powerful technologies benefit society globally while reasonably managing emerging risks.
-> ensure these powerful technologies remain safe so that their benefits to global society can be realized.

a measure of speculation about potential trajectories -> reasoned analysis of potential trajectories

p6
"enable a policy environment that promotes innovation" - sure, but does not imply deregulation. For example, absence of liability does not promote innovation around safety and accuracy, which are key areas where current AI systems are lacking.

"Experts disagree on the probability of these risks." - does it really matter if it’s 10% or 30%? And what do those extinction risks refer to? Probability of nearly immediate extinction after the introduction of uncontrolled AGI? What about the long-term risks from, say, 100 years of uncontrolled AGI? Would that not approach 100%, if it’s 10% in the immediate aftermath?

"whether it is possible to build" – I’m not sure who the experts are who have a serious argument showing it is impossible to build. I’ve not seen a single convincing technical argument. If you have, please cite it!

"policy activity on foundation models includes work by" Singapore?

p7

"to prevent algorithmic decision-making" - needs qualifying? Surely not a blanket ban on algorithms, period.

evidence-generating mechanisms to provide "stronger evidence of impending risk" would mean … deliberately causing those risks? Please spell out the nature of the evidence gap these mechanisms fill

"internalize societal externalities" - surely liability is the key here? Insurance typically follows

"carefully tailored policy ensures and accelerates the development of robust scientific understanding" - yes, especially if policy requires quantifiable evidence of safety (cf nuclear power)

p8

"The model achieved gold medal performance without any custom coding-specific test-time strategies defined by humans." - did its training set include previous Olympiads and other CS tests?

"Increased scale has largely, but not exclusively, driven recent improvements" - I think this is a little bit outdated: many/most observers think scaling is over

p9

"Collectively, the current, inconclusive level of evidence for these risks motivates our exploration of evidence- generating policy mechanisms [9, 21]."
Does the report really engage with the "pitfalls" of "evidence-based policy" in ref. 21??

"comes from inference scaling" - that seems an odd way of putting it. It comes from doing inference, which previously LLMs were not doing, and a lot of work on how to make inference effective by choosing what inference steps to make. Quite analogous to early chess programs. "Inference scaling" for chess programs set in during the 1970s, once the basic paradigm settled down with alpha-beta pruning etc. I don’t think the LLM inference paradigm has settled down.

"paradigm shift toward AI agents." - probably need to explain a bit more. What does this mean? I think it’s a mistake to think that ordinary LLMs are not agents - they emit speech acts, just like humans, that affect the world.

p11 "Many have published safety frameworks articulating thresholds that, if passed, will trigger concrete safety-focused actions." - And several CEOs of major AI companies have stated that human extinction is a significant risk from their technology; no analogous admission has been forthcoming from oil and tobacco companies.

p13
"policymaking had focused on responding to individual incidents rather than implementing structural security reforms, eventually resulting in substantial real-world harms." - worth noting that the computer security industry has no financial interest in "structural security reforms", which would put it out of business.

"costs the United States millions of dollars annually" - 2024 US GDP is about $23.5 trillion; so 0.9% to 4.1% is $211B to $963B annually, which is a lot more than "millions". It would be accurate to say "hundreds of billions" (and that’s just consumers; industry ransomware attacks, IP theft, etc etc add possibly trillions more).

"early design choices and security protocols will shape long-term governance challenges." - Yes, but it’s far from clear what "design choices" are being made now and locked in. These could come soon in terms of, say, interaction protocols between AI agents, rules about retention of interaction records, etc etc.

"It also reveals" - not sure what "It" refers to here.

p14
"these roles can be constrained both by internal divisions within these agencies" - not sure what is being referred to here; as written it sounds like a generic and unnecessary point that could apply to any agency around any issue.

Re "Policy windows do no remain open indefinitely" - here is Paul Berg, convener of the 1975 Asilomar Workshop, on that question. Worth including this quote, I think:

“There is a lesson in Asilomar for all of science: the best way to respond to concerns created by emerging knowledge or early-stage technologies is for scientists from publicly funded institutions to find common cause with the wider public about the best way to regulate — as early as possible. Once scientists from corporations begin to dominate the research enterprise, it will simply be too late.”
-- Paul Berg, “Asilomar 1975: DNA modification secured,” Nature 455 (2008): 290–91.

"If those who speculate about the most extreme risks are right" - the word "speculate" here seems pejorative and inconsistent with previous arguments in the report about the value of predictive analysis. How about "If those whose analysis points to the most extreme risks are right" "and we are uncertain if they will be" is unnecessary given the conditional "If..."

"considerable latitude to make decisions" - considerable latitude to make *informed* decisions. It's often the information asymmetry that leads to abuses, addiction, etc.

p15 "between $200 and $240 billion" - sounds like a lot, doesn't it?
Global early deaths in the 20th century from tobacco are estimated at roughly 100 million, and the current rate is 8 million deaths per year.
Just for the 20th century, that's only $2000 to $2400 per death, whereas a typical wrongful death settlement would typically be 100x to 1000x higher than that.
Put another way, the damage runs into hundreds of trillions of dollars. I feel the report softpedals on this and similar issues.

p17 "In assessing future trajectories for AI..." - good to see these xamples being cited.

p19 "A False Safety-Innovation Binary" - perhaps this point should be part of the executive summary. The industry mantra "regulation stifles innovation" and even the notion of a risk-benefit tradeoff have dominated the discussion for five decades.

2.6 fails to learn the lesson stated by Paul Berg. In the case of the internet, tobacco, and climate change, corporate interests crushed academic findings about risk, resulting in damages ranging from trillions to quadrillions of dollars.