I would like to answer question 2.a specifically:

> *What, if any, are the risks associated with widely available model weights? How do these risks change, if at all, when the training data or source code associated with fine tuning, pretraining, or deploying a model is simultaneously widely available?*

Credentials: I am an American citizen living abroad. I work closely with the Future of Humanity Institute at Oxford, the blog lesswrong.com, and the online “AI safety fundamentals” course by BlueDotAI. All three of these entities were created to respond to the existential threat posed by AI. All of us stand by the statement signed, among others, by the CEOs of all the major AI labs (including OpenAI): “*Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.”*

Thesis: Open-sourcing model weights over 10 billion parameters should not be legal, because it drastically increases the chance of human extinction by AI. Open-sourcing model weights puts me, my family, and your children at risk of death.

Argument: Sam Altman (CEO of OpenAI), Dario Amodei (CEO of Anthropic), and Demis Hassabis (CEO of DeepMind) ***all*** believe that Artificial General Intelligence (AGI), or superhuman intelligence, is less than 5 years away.

This is alarming news because we have not yet solved the “control problem” or “alignment problem”, where an AI model does exactly as you say. This has had no obvious negative consequences so far, because you can always just fine-tune a model a little more if it makes mistakes.

But when it comes to AGI, which is vastly smarter than current models, failing to solve the control problem could have catastrophic outcomes. The two top-cited AI researchers in the world (Geoffrey Hinton and Yoshua Bengio) believe that if we reach AGI before we solve the control problem, human extinction will happen *by default*.

Publishing model weights is just like publishing the blueprints to thermonuclear weapons, only worse. If anyone, anywhere, can build AGI with more ease than before, then human extinction becomes more probable. Computing and data are still becoming exponentially cheaper, and if *on top of that* Meta and Mistral (two prominent AI labs) are openly publishing their weights, then building a world-destroying machine becomes exponentially easier. Every year, building machines of awesome power becomes easier, while safety work is left in the dust as it lags along.

Scott Alexander, a renowned blogger, has made the point more eloquently and precisely than almost anyone else. The following article has been read (I know for a fact) by many US officials. It is short, and I highly recommend it:

slatestarcodex.com/2015/12/17/should-ai-be-open/

This excerpt talks about the control problem:

> *OpenAI's strategy also skips over a second aspect of AI risk: the control problem.*
>
> *All of this talk of "will big corporations use AI?" or "will Dr. Evil use AI?" or "Will AI be used for the good of all?" presuppose that you can use an AI. You can certainly use an AI like the ones in chess-playing computers, but nobody's very scared of the AIs in chess-playing computers either. What about AIs powerful enough to be scary?*
>
> *Remember the classic programmers' complaint: computers always do what you tell them to do instead of what you meant for them to do. Computer programs rarely do what you want the first time you test them. Google Maps has a relatively simple task (plot routes between Point A and Point B), has been perfected over the course of years by the finest engineers at Google, has been 'playtested' by tens of millions of people day after day, and still occasionally does awful things like suggest you drive over the edge of a deadly cliff, or tell you to walk across an ocean and back for no reason on your way to the corner store.*
>
> *…*
>
> *That means a serious risk of superhuman AIs that want to do the equivalent of hurl us off cliffs, and which are very resistant to us removing that desire from them. We may be able to prevent this, but it would require a lot of deep thought and a lot of careful testing and prodding at the cow-level AIs to make sure they are as prepared as possible for the transition to superhumanity.*
>
> ***And we lose that option by making the AI open source.*** *Make such a program universally available, and while Dr. Good is busy testing and prodding, Dr. Amoral has already downloaded the program, flipped the switch, and away we go.*

Making model weights open-source will increase the chance that your children get slaughtered at the hands of a superhuman AI within the next 5 years. Hinton and Bengio, who are known as the "godfathers of AI", give human extinction a 20% probability of occurring. **That's worse odds for your children than Russian roulette.** I hate being that frank, but I must. There's no time to dilly-dally.

…

The more information is widely available, whether it be just the model weights or the model weights *and* the training data and source code, the worse this risk becomes. Making model weights open-source will drastically increase the chance of human extinction.