Classification	Hardware trigger	Capabilities trigger *(1-3 based on Anthropic's classification)*	Requirements for training	Requirements for deployment
ASL-1	None	**Negligible potential for harm:** Systems which pose no meaningful catastrophic risk, for example a 2018 LLM or an AI system that only plays chess.	None	None
ASL-2	None	**Potential for minor harm:** Systems that show early signs of dangerous capabilities – for example ability to give instructions on how to build bioweapons – but where the information is not yet useful due to insufficient reliability or not providing information that e.g. a search engine couldn't.	Registration with national authority with juristiction over the lab	**Safety audits** by national authorities wherever the model can be used, including blackbox and whitebox red-teaming
ASL-3	100 yotta-FLOP (10^{26})	**Potential for major harm:** Systems that substantially increase the risk of catastrophic misuse compared to non-AI baselines (e.g. search engines or textbooks) OR that show low-level autonomous capabilities, alone or in combination with other available techniques OR that classify as "very capable foundation models" under provisions of the EU AI act.	Pre-approval of safety plan by national authority with juristiction over the lab	**Quantitative safety bounds:** National authorities wherever the model can be used must approve lab-submitted assesment bounding risk of major harm below authorized levels.
ASL-4	ronnaFLOP (10^{27})	**AGI potential:** Systems with potential to enable AGI alone or in combination with other available techniques, where AGI is defined as AI capable of performing all economically valuable cognititive tasks at human expert level.	Pre-approval of safety plan by IAIA	**Provable safety:** The IAIA must certify lab-submitted formal verification that the model provably meets required specifications, including cybersecurity, controllability, a non-removable kill-switch, alignment with human values, and robustness to malicious use.