“Levels of defense” in AI safety