Une image contenant conception, art

Description générée automatiquement avec une confiance moyenne**Leader** Apr 20th 2023 Adapted from THE ECONOMIST

**How to worry wisely about artificial intelligence**

*Rapid progress in AI is arousing fear as well as excitement. How worried should you be?*

**§1** “Should we automate away all the jobs, including the fulfilling ones? Should we develop non-human minds that might eventually outnumber, outsmart...and replace us? Should we risk loss of control of our civilisation?” These questions were asked last month in an open letter from the Future of Life Institute, an NGO. It called for a six-month “pause” in the creation of the most advanced forms of artificial intelligence (AI), and was signed by tech luminaries including Elon Musk. It is the most prominent example yet of how rapid progress in AI has sparked anxiety about the potential dangers of the technology.

**§2** In particular, new “large language models” (LLMs)—the sort that powers [ChatGPT](safari-reader://www.economist.com/graphic-detail/2023/04/14/chatgpt-could-replace-telemarketers-teachers-and-traders), a chatbot made by OpenAI, a startup—have surprised even their creators with their unexpected talents as they have been scaled up. Such “emergent” abilities include everything from solving logic puzzles and writing computer code to identifying films from plot summaries written in emoji.

**§3** Proponents of AI argue for its potential to solve big problems by developing new drugs, designing new materials to help fight climate change, or untangling the complexities of fusion power. To others, the fact that AIs’ capabilities are already outrunning their creators’ understanding risks bringing to life the science-fiction disaster scenario of the machine that outsmarts its inventor, often with fatal consequences.

**§4** This bubbling mixture of excitement and fear makes it hard to weigh the opportunities and risks. But lessons can be learned from other industries, and from past technological shifts. So what has changed to make AI so much more capable? How scared should you be? And what should governments do?

**§5** The first wave of modern AI systems, which emerged a decade ago, relied on carefully labelled training data. Once exposed to a sufficient number of labelled examples, they could learn to do things like recognise images or transcribe speech. Today’s systems do not require pre-labelling, and as a result can be trained using much larger data sets taken from online sources. LLMs can, in effect, be trained on the entire internet—which explains their capabilities, good and bad.

**§6** Those capabilities became apparent to a wider public when ChatGPT was released in November. A million people had used it within a week; 100m within two months. It was soon being used to generate school essays and wedding speeches. ChatGPT’s popularity prompted rival firms to release chatbots too.

**§7** Some of these produced strange results. LLMs produce answers that have the patina of truth, but often contain factual errors or outright fabrications. Even so, Microsoft, Google and other tech firms have begun to incorporate LLMs into their products, to help users create documents and perform other tasks.

**§8** The recent acceleration in both the power and visibility of AI systems, and growing awareness of their abilities and defects, have raised fears that the technology is now advancing so quickly that it cannot be safely controlled. Hence the call for a pause, and growing concern that AI could threaten not just jobs, factual accuracy and reputations, but the existence of humanity itself.

**§9** The fear that machines will steal jobs is centuries old. But so far new technology has created new jobs to replace the ones it has destroyed. Machines tend to be able to perform some tasks, not others, increasing demand for people who can do the jobs machines cannot. Could this time be different? A sudden dislocation in job markets cannot be ruled out, even if so far there is [no sign of one](safari-reader://www.economist.com/business/2023/04/19/how-businesses-are-experimenting-with-chatgpt-like-services). Previous technology has tended to replace unskilled tasks, but LLMs can perform some white-collar tasks, such as summarising documents and writing code.

**§10** The degree of existential risk posed by AI has been hotly debated. Experts are divided. In a survey of AI researchers carried out in 2022, 48% thought there was at least a 10% chance that AI’s impact would be “extremely bad (eg, human extinction)”. But 25% said the risk was 0%; the median researcher put the risk at 5%. The nightmare is that an advanced AI causes harm on a massive scale, by making poisons or viruses, or persuading humans to commit terrorist acts. It need not have evil intent: researchers worry that future AIs may have goals that do not align with those of their human creators.

**§11** Such scenarios should not be dismissed. But all involve a huge amount of guesswork, and a leap from today’s technology. Moreover, experts tend to overstate the risks in their area, compared with other forecasters. (And Mr Musk, who is launching his own AI startup, has an interest in his rivals downing tools.) Imposing heavy regulation, or indeed a pause, today seems an over-reaction. A pause would also be unenforceable.

**§12** Regulation is needed, but for more mundane reasons than saving humanity. Existing AI systems raise real concerns about bias, privacy and intellectual-property rights. The key is to balance the promise of AI with an assessment of the risks, and to be ready to adapt.

**§13** So far governments are taking three different approaches. At one end of the spectrum is Britain (and also the USA), which has proposed a “light-touch” approach with no new rules or regulatory bodies, but applies existing regulations to AI systems. The aim is to boost investment and turn Britain into an “AI superpower”.

**§14** The EU is taking a tougher line. Its proposed law categorises different uses of AI by the degree of risk, and requires increasingly stringent monitoring and disclosure as the degree of risk rises from, say, music-recommendation to self-driving cars. Some uses of AI are banned altogether, such as subliminal advertising and remote biometrics. Firms that break the rules will be fined. For some critics, these regulations are too stifling.

**§15** But others say an even sterner approach is needed. Governments should treat AI like medicines, with a dedicated regulator, strict testing and pre-approval before public release. China is doing some of this, requiring firms to register AI products and undergo a security review before release. But safety may be less of a motive than politics: a key [requirement](safari-reader://www.economist.com/china/2023/04/18/can-xi-jinping-control-ai-without-crushing-it) is that AIs’ output reflects the “core value of socialism”.

**§16** What to do? The light-touch approach is unlikely to be enough. If AI is as important a technology as cars, planes and medicines, then, like them, it will need new rules. Accordingly, the EU’s model is closest to the mark, though its classification system is not flexible enough. Compelling disclosure about how systems are trained, how they operate and how they are monitored, and requiring inspections, would be comparable to similar rules in other industries.

**§17** This could allow for tighter regulation over time, if needed. A dedicated regulator may then seem appropriate; so too may intergovernmental treaties, similar to those that govern nuclear weapons, should plausible evidence emerge of existential risk. To monitor that risk, governments could form a body modelled on CERN, a particle-physics laboratory, that could also study AI safety and ethics—areas where companies lack incentives to invest as much as society might wish.

**§18** This powerful technology poses new risks, but also offers extraordinary opportunities. Balancing the two means treading carefully. A measured approach today can provide the foundations on which further rules can be added in future. But the time to start building those foundations is now.