Philosopher Nick Bostrom believes it’s entirely possible that artificial intelligence (AI) could lead to the extinction of *Homo sapiens*. In his 2014 bestseller *Superintelligence: Paths, Dangers, Strategies*, Bostrom paints a dark scenario in which researchers create a machine capable of steadily improving itself. At some point, it learns to make money from online transactions and begins purchasing goods and services in the real world. Using mail-ordered DNA, it builds simple nanosystems that in turn create more complex systems, giving it ever more power to shape the world.

Now suppose the AI suspects that humans might interfere with its plans, writes Bostrom, who’s at the University of Oxford in the United Kingdom. It could decide to build tiny weapons and distribute them around the world covertly. “At a pre-set time, nanofactories producing nerve gas or target-seeking mosquito-like robots might then burgeon forth simultaneously from every square meter of the globe.”

For Bostrom and a number of other scientists and philosophers, such scenarios are more than science fiction. They’re studying which technological advances pose “existential risks” that could wipe out humanity or at least end civilization as we know it—and what could be done to stop them. “Think of what we’re trying to do as providing a scientific red team for the things that could threaten our species,” says philosopher Huw Price, who heads the Centre for the Study of Existential Risk (CSER) here at the University of Cambridge.

The idea of science eliminating the human race can be traced all the way back to *Frankenstein*. In Mary Shelley’s novel, the monster gets angry at his creator, Victor Frankenstein, for having spurned him. He kills Frankenstein’s little brother William, but then offers the doctor a deal: Make a female companion for me and we will leave you in peace and go to South America to live out our days. Frankenstein starts working on the bride, but realizes that the couple might reproduce and outcompete humans: “A race of devils would be propagated upon the earth who might make the very existence of man a condition precarious and full of terror.” He destroys the half-finished female, reigniting the creature’s wrath and bringing about his own demise.

“I think *Frankenstein* illustrates the point beautifully,” says physicist Max Tegmark of the Massachusetts Institute of Technology (MIT) in Cambridge, a board member of CSER and a co-founder of a similar think tank, the Future of Life Institute (FLI), near MIT. “We humans gradually develop evermore-powerful technology, and the more powerful the tech becomes, the more careful we have to be, so we don’t screw up with it.”

The study of existential risks is still a tiny field, with at most a few dozen people at three centers. Not everyone is convinced it’s a serious academic discipline. Most civilization-ending scenarios—which include humanmade pathogens, armies of nanobots, or even the idea that our world is a simulation that might be switched off—are wildly unlikely, says Joyce Tait, who studies regulatory issues in the life sciences at the Innogen Institute in Edinburgh. The only true existential threat, she says, is a familiar one: a global nuclear war. Otherwise, “There is nothing on the horizon.”

Harvard University psychologist Steven Pinker calls existential risks a “useless category” and warns that “Frankensteinian fantasies” could distract from real, solvable threats such as climate change and nuclear war. “Sowing fear about hypothetical disasters, far from safeguarding the future of humanity, can endanger it,” he writes in his upcoming book *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*.

But advocates predict the field will only get more important as scientific and technological progress accelerates. As Bostrom pointed out in one paper, much more research has been done on dung beetles or *Star Trek* than on the risks of human extinction. “There is a very good case for saying that science has basically ignored” the issue, Price says. HUMANITY HAS ALWAYS FACED the possibility of an untimely end. Another asteroid of the size that ended the dinosaurs’ reign...
A nuclear bomb is detonated at the U.S. Pacific Proving Grounds in 1958. Global nuclear warfare is one of the "existential risks" that could end civilization as we know it.
A glossary of Frankenwords
Along with fears, the Frankenstein story has inspired hundreds of whimsical names for products and phenomena. Here’s a selection; you’ll find more at http://scim.ag/Frankenwords.

By Jon Cohen

**Frankenbrooms**
Supercoated broomheads used in curling, a winter sport played on ice.

**Frankencell**
J. Craig Venter’s attempt to create an artificial cell containing the smallest possible number of essential genes.

**Frankenfears**
Exaggerated concerns about transgenic food (Frankenfood) and other products of genetic engineering.

**Frankenforests**
Engineered trees that grow more quickly, absorbing more carbon dioxide and providing more wood and pulp without the need for toxic chemicals.

**Frankengene**
The catechol-O-methyltransferase gene. In 2008, researchers linked variants in the gene to the strong, frightful reaction some people have to horror movies.

**Frankenmoths**
A male diamondback moth engineered to spread a lethal gene to females, creating nonviable offspring that reduce the moth’s toll on crops. Cousins include Frankenflies and Frankenmosquitoes.

**Frankenmouse**
A genetically or surgically altered mouse. Variations have included the “oncomouse” that’s prone to cancer and the “earmouse” that had a human-shaped ear (photo, left).

**Frankenmums**
Mothers who freeze eggs for their infertile daughters to use.

**Frankenpets**
Transgenic dogs that would repel fleas; cats that would not cause allergies.

**Frankenpines**
Cellphone towers that resemble pine trees.

**Frankenbrooms**

**Frankencell**

**Frankenfears**

**Frankenforests**

**Frankengene**

**Frankenmoths**

**Frankenmouse**

**Frankenmums**

**Frankenpets**

**Frankenpines**

Church says a “crunch,” in which a large part of the world population dies, is more likely than a complete wipe-out. “You don’t have to turn the entire planet into atoms,” he says. Disrupting electrical grids and other services on a huge scale or releasing a deadly pathogen could create chaos, topple governments, and send humanity into a downward spiral. “You end up with a medieval level of culture,” Church says. “To me that is the end of humanity.”

Existential risks stemming from the life sciences are perhaps easiest to imagine. Pathogens have proved capable of killing off entire species, such as the frogs that have fallen victim to the amphibian fungus *Batrachochytrium dendrobatidis*. And four influenza pandemics have swept the world in the past century, including one that killed up to 50 million people in 1918 and 1919. Researchers are already engineering pathogens that in principle could be even more dangerous. Worries about studies that made the H5N1 bird flu strain more easily transmissible between mammals led the United States to halt such research until late last year. Terrorists or rogue states could use labmade agents as a weapon, or an engineered plague could be released accidentally.

Rees has publicly wagered that by 2020, “bioterror or bioerror will lead to 1 million casualties in a single event.” Harvard microbiologist Marc Lipsitch has calculated that the likelihood of a labmade flu virus leading to an accidental pandemic is between one in 1000 and one in 10,000 per year of research in one laboratory; Ron Fouchier of Erasmus MC in Rotterdam, the Nether-
The release of a dangerous pathogen might cause a “crunch” in the human population.

lands, one of the researchers involved in the H5N1 studies, has dismissed that estimate, saying the real risk is more like one in 33 billion per year and lab.

One measure against “bioerror” might be to make researchers who carry out risky experiments buy insurance; that would require an independent assessment of the risk and would force researchers to face up to it, Lipsitch says. Still, the most important countermeasure is to strengthen the world’s capacity to contain an outbreak early on, he adds, for instance with vaccines. “For biological risks, short of a really massive, coordinated, parallel attack around the world, the only way we are going to get to a really catastrophic scenario is by failing to control a smaller scenario,” he says.

**VIRUSES ARE UNLIKELY** to kill every last human, Bostrom says; for him and others, it is AI that poses truly existential threats. Most scenarios center on machines outsmarting humans, a feat called “superintelligence.” If such AI were ever achieved and it acquired a will of its own, it might turn malevolent and actively seek to destroy humans, like HAL, the computer that goes rogue aboard a spaceship in Stanley Kubrick’s film *2001: A Space Odyssey*.

Most AI experts worry less about machines rising up to overthrow their creators, however; than about them making a fatal mistake. To Tallinn, the most plausible way in which AI could end humanity is if it simply pursued its goals and, along the way, heedlessly created an environment fatal to humans. “Imagine a situation where the temperature rises by 100° or is lowered by 100°. We’d go extinct in a matter of minutes,” Tallinn says. Tegmark agrees: “The real problem with AI is not malice, it’s incompetence,” he says.

A current-day analogy is the 2015 tragedy in which a suicidal Germanwings pilot told his plane’s computer to descend to an altitude of 100 meters while flying over the French Alps. The machine complied, killing all 150 on board, even though it had GPS and a topographic map. “It did not have a clue about even the simplest human goal,” Tegmark says. To avoid such calamities, scientists are trying to figure out how to teach AI human values and make sure they stick, a problem called “value alignment.” “There might be fewer than 20 people who work full time on technical AI safety research,” Bostrom says. “A few more talented people might substantially increase the rate of progress.”

**CRITICS SAY THESE EFFORTS** are unlikely to be useful, because future threats are inherently unpredictable. Predictions were a problem in every “foresight exercise” Tait has taken part in, she says. “We’re just not good at it.” Even if you foresee a risk, economic, political, and societal circumstances will all affect how it plays out. “Unless you know not only what is going to happen, but how it is going to happen, the information is not much use in terms of doing something about it,” Tait says.

Pinker thinks the scenarios reveal more about human obsessions than real risks. We are drawn to prospects “that are highly improbable while having big impacts on our fitness, such as illicit sex, violent death, and Walter-Mittyish feats of glory,” he writes. “Apocalyptic storylines are undoubtedly gripping—they are a supernormal stimulus for our morbid obsessions.” Sure, he says, one can imagine a malevolent, powerful AI that people can no longer control. “The way to deal with this threat is straightforward: Don’t build one.”

Tallinn argues it’s better to be safe than sorry. A 2017 survey showed that 34% of AI experts believed the risks associated with their work are an important problem; 5% said they are “one of the most important problems.” “Imagine you’re on a plane, and 40% of experts think that there is a bomb on this plane,” Tallinn says. “You’re not going to wait for the remaining experts to be convinced.”

Price says that critics who accuse him and his colleagues of indulging in science fiction are not entirely wrong: Producing doomsday scenarios is not that different from what Shelley did. “The first step is to imagine that range of possibilities, and at that point, the kind of imagination that is used in science fiction and other forms of literature and film is likely to be extremely important,” he says.

Scientists have an obligation to be involved, says Tegmark, because the risks are unlike any the world has faced before. Every time new technologies emerged in the past, he points out, humanity waited until their risks were apparent before learning to curtail them. Fire killed people and destroyed cities, so humans invented fire extinguishers and flame retardants. With automobiles came traffic deaths—and then seat belts and airbags. “Humanity’s strategy is to learn from mistakes,” Tegmark says. “When the end of the world is at stake, that is a terrible strategy.”
Taming the monsters of tomorrow
Kai Kupferschmidt

Science 359 (6372), 152-155.
DOI: 10.1126/science.359.6372.152

Use of this article is subject to the Terms of Service