

The science fiction science method

<https://doi.org/10.1038/s41586-025-09194-6>

Received: 19 November 2024

Accepted: 23 May 2025

Published online: 6 August 2025

 Check for updates

Iyad Rahwan^{1✉}, Azim Shariff^{2✉} & Jean-François Bonnefon^{3✉}

Predicting the social and behavioural impact of future technologies before they are achieved would enable us to guide their development and regulation before these impacts get entrenched. Traditionally, this prediction has relied on qualitative, narrative methods. Here we describe a method that uses experimental methods to simulate future technologies and collect quantitative measures of the attitudes and behaviours of participants assigned to controlled variations of the future. We call this method ‘science fiction science’. We suggest that the reason that this method has not been fully embraced yet, despite its potential benefits, is that experimental scientists may be reluctant to engage in work that faces such serious validity threats. To address these threats, we consider possible constraints on the types of technology that science fiction science may study, as well as the unconventional, immersive methods that it may require. We seek to provide perspective on the reasons why this method has been marginalized for so long, the benefits it would bring if it could be built on strong yet unusual methods, and how we can normalize these methods to help the diverse community of science fiction scientists to engage in a virtuous cycle of validity improvement.

Imagine that behavioural scientists managed to predict the effect of social media on mental health and democratic life before social media actually existed. That is, imagine that they had designed simulations of what they thought social media might look like, and recorded participants’ interactions with this speculative technology. They might have noted a tendency for participants to focus on upward social comparisons, leading to self-esteem issues¹, or a focus on moral outrage, resulting in exaggerated feelings of polarization². These speculative findings could have guided us in designing and regulating social media with the benefit of foresight, instead of constantly playing catch-up with their effects³.

Social media is just one example of a technology that could have benefited from early, speculative behavioural research before widespread deployment. For example, when genetically modified foods were taken to market in the 1990s, they faced a strong public opposition that caught both the industry and policymakers off guard. Companies and regulatory bodies had not anticipated the psychological concerns of the public about this technology⁴, which led to a persistent cycle of mistrust that has not been fully resolved even 30 years later⁵. Today, many people still manifest absolute moral opposition to genetically modified foods⁶, even though they understand very little about them⁷. It is of course impossible to claim that history would have been different if behavioural scientists had conducted experiments on the acceptability of genetically modified foods before the technology was fully developed. However, the experience with genetically modified foods provided a valuable lesson, leading to the normalization of prospective social acceptability studies in other fields. For example, people were polled about their reactions to some potential applications of nanotechnologies well before these applications became possible^{8,9}.

In this Perspective, we first observe that despite a clear and consensual need to predict the social and behavioural impact of future

technologies, it remains uncommon to do so with the experimental methods and quantitative measures that are typical of contemporary behavioural science. Consequently, we advocate for what we term ‘science fiction science’ (sci-fi-sci): the application of the scientific method to the science fiction project of anticipating behavioural and social changes driven by a speculative technology. After providing a more detailed definition of sci-fi-sci, we review past and current examples, illustrating its primary challenges: prospective validity and cost–benefit analysis. We discuss the principles and methods that have and could be used to overcome these challenges. Our goals are to synthesize the reasons why we need sci-fi-sci, and the reasons why it is so disorganized still; to foster a community of researchers under this unified banner; to legitimize the unconventional methods required for the unconventional enterprise of conducting behavioural experiments centred on technologies that do not yet exist; and to help the diverse community of science fiction scientists to engage in a virtuous cycle of methodological improvement and community growth.

Technology-driven futures

New technologies transform societies by changing what people can do, what they actually do, and what they think is acceptable to do¹⁰. For example, the introduction of a reliable birth control pill gave women unprecedented control over their reproductive lives, education and careers, transforming social norms around premarital sex, family planning and women’s participation in the workforce^{11,12}. In parallel, the technology that made organ donation easier by maintaining heart and lung function in patients declared brain dead changed the way doctors and patients think of the ethics of ending life¹³. Yet other technologies transformed society because they were not accompanied by adequate changes in behaviour and social norms. For example, sophisticated

¹Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany. ²Department of Psychology, University of British Columbia, Vancouver, British Columbia, Canada.

³Toulouse School of Economics, CNRS (TSM-R), University of Toulouse-Capitole, Toulouse, France. ✉e-mail: rahwan@mpib-berlin.mpg.de; shariff@psych.ubc.ca; jean-francois.bonnefon@tse-fr.eu

Box 1

Examples of policy debates about nascent or speculative technologies

Connected and autonomous vehicles. Autonomous driving technology is still being tested and fully autonomous cars are confined to specific cities. Yet, a debate ensued early on about the ethical implications of intelligent machines making life and death decisions in unavoidable accident scenarios⁶⁶. These debates have led to significant mobilization by civil society⁶⁷ and specific policy recommendations²⁶, some of which have informed legislation⁶⁸.

Social credit systems. Governments do not yet have the capacity to deploy AI systems that would monitor every behaviour of all their citizens in real time to calculate and release social scores, but the behavioural, social and political impact of these hypothetical systems is already the object of much speculation^{69–72}, to the point where the European Union is considering a pre-emptive ban of this technology for its member governments⁷³.

Embryo screening. Fertility clinics are only in the very early stages of providing parents with the option to choose or reject specific genetic traits in their offspring⁷⁴. However, bioethicists are already considering the ethical trade-offs, social dilemmas and collective action problems that may arise if parents gain the ability to selectively determine cognitive, moral, physical and immune traits^{75–77}.

Ectogenesis. Medical research has not yet achieved ectogenesis (artificial gestation of human fetuses), but the field of speculative bioethics is already debating whether this potential technology might end gender-based oppression or exacerbate other inequalities, and to what extent it could disrupt social and legal norms around motherhood and employment rights^{78–81}.

tracking technologies enabled tech giants to accumulate vast amounts of personal data, leading to a commodification of human experience because people failed to adjust their behaviours and norms in a way that would have sustained privacy^{14,15}.

In all these examples, we have the benefit of hindsight about the impact of each technology. However, our capacity to now alter these impacts—if we wanted to—has become limited by behaviours, norms and institutions that are now deeply entrenched. By contrast, we have much power to steer and regulate technologies that are in their early or speculative stage—but when a technology is at this speculative stage, we know little about the impact it may have on society, or how our actions may change this impact (Box 1). This tension is known as the Collingridge control dilemma¹⁶: the moment when we have the best chance to shape the social impact of a technology is also the moment when we know the least about what it will actually do and how. One way out of this dilemma is to keep regulation as agile and flexible as possible so that it can quickly adapt to our emerging understanding of a technology. Another escape from the dilemma is to attempt to predict, as early as possible, the different futures that may unfold under the influence of a given technology—for example, what futures are the most likely if the technology is left unregulated, what regulations would be the most acceptable to the people of the future, and what futures may further unfold as a result of these regulations.

Anticipating technology-driven futures has long been the project of futures studies, which primarily use qualitative methods such as Delphi surveys and other techniques aimed at organizing communication between expert panellists¹⁷. Future studies typically attempt

to compose various scenarios that describe the future on the basis of expert intuition, and then work backwards to imagine what actions or events may lead to these scenarios¹⁸. Besides futurists, ethicists and legal scholars also show interest in the social and moral impact of speculative technologies, and they also primarily use qualitative methods, such as normative analysis and case studies^{19,20}. What is much less common is for scientists to weigh in on these debates by producing quantitative data extracted from experimental methods and behavioural measures²¹. These methods are routinely used to examine the impact of existing technologies, but their presence sharply diminishes when turning to future technologies. In other words, there seems to be no well-established field that would be the quantitative, experimental, behavioural counterpart to futures studies and other qualitative explorations of the impact of speculative technologies.

Science fiction science

Let us summarize the problem so far. We start from a technology that is not yet available, but is likely to become available in the future, with some ill-defined capabilities and limitations. We believe that it may transform individual behaviours, social norms and institutions, with a mixture of desirable and undesirable effects. The potential magnitude of these effects is sufficient for us to worry about them even before the technology is achieved, in the hope that we can steer its development towards the most desirable effects, and prepare to regulate against its remaining undesirable effects. The challenge is that because the technology does not exist yet, we have no behavioural data about what people will think and do when it becomes available, no controlled experiments and no tentative quantification of its positive and negative effects.

To collect such data, we need to develop methods and principles for a speculative behavioural science. We call this method sci-fi-sci because, at its core, it consists of applying the scientific method to the science fiction project. Indeed, science fiction writers provide elaborate thought experiments about the social and behavioural impact of future technologies. As science fiction author Frederik Pohl once wrote, paraphrasing Isaac Asimov: ‘Somebody once said that a good science fiction story should be able to predict not the automobile but the traffic jam’²². Sci-fi-sci is an attempt to turn the thought experiments of science fiction into behavioural experiments; science fiction tells the stories of future humans, but sci-fi-sci attempts to study them in the laboratory.

More precisely, sci-fi-sci is the application of the scientific method to anticipate and study the behavioural, psychological, attitudinal and social impacts of speculative technologies that do not yet exist. By recruiting present-day participants and immersing them into controlled experimental simulations of possible futures, sci-fi-sci obtains quantitative measures of their thoughts, attitudes and behaviours. This approach utilizes rigorous experimental methods—including control and treatment conditions with independent and orthogonal variable manipulation—to not only survey attitudes but also measure direct behaviours as participants interact with simulated future technologies. By doing so, sci-fi-sci aims to uncover potential sociotechnical benefits and harms, and to offer insights into how different factors (including marketing) can influence public perception and behaviour. It serves as the quantitative, experimental and behavioural counterpart to futures studies and other qualitative explorations of the impact of speculative technologies.

Our own research provides a notable example of sci-fi-sci (see Box 2 for further details). Between 2016 and 2018, we published several articles on fully autonomous vehicles (AVs), a technology that was widely anticipated, but had not yet been achieved. In these articles, we surveyed participants on the ethical preferences they would want these vehicles to follow in the event of an unavoidable collision, where the vehicle must decide which road users to spare and which to sacrifice^{23–25}.

Box 2

Case study A: ethical dilemmas of autonomous vehicles

Period of sci-fi-sci studies: 2016 to present day.

Sci-fi-sci question. How do citizens and consumers wish AVs to prioritize the safety of different road users in unavoidable accidents? Technology ethicists⁸² and transportation experts⁶⁶ were already discussing such possibilities before the mid-2010s.

Technological plausibility. Many proof-of-concept tests took place through the US Defense Advanced Research Projects Agency Urban Challenge in 2007 (TRL=6). By 2015, various US states even allowed AV testing on public roads (TRL=7–8).

Temporal proximity. Spurred by early success, the market invested billions of dollars into AV technology development, suggesting that the technology was very proximal, even while estimates of its readiness varied. Indeed, by 2024, Waymo had deployed its fully autonomous taxi service in the City of San Francisco (TRL=9–10).

Magnitude of effect. AVs could have substantial socioeconomic effects such as changing urban land use and altering commuter behaviour or even the value of time⁸³. Many of these outcomes, often based on computational simulation, are highly sensitive to modelling assumptions⁸⁴, and are thus difficult to study behaviourally. By contrast, it is more feasible to study consumer attitudes that may shape the early adoption of AVs, and how consumers and citizens may react to different design features and regulatory regimes.

Sample of sci-fi-sci studies. Starting in 2016, a series of studies attempted to anticipate how people would react to AV accident dilemmas. Initial studies used text vignettes²³ that highlighted the disconnect between the preferences of citizens (AVs should save as many lives as possible) and those of consumers (AVs should prioritize passengers). The Moral Machine experiment²⁵ used a conjoint design with visualizations of binary choices, crowdsourcing more than 40 million decisions from people worldwide, and highlighting cross-cultural differences. More recent studies used virtual reality to situate participants in the expected reality of a driverless car carrying out complex moral decisions^{49–51}.

Subsequent relevance. By engaging millions of citizens worldwide, the Moral Machine experiment itself has contributed to a wide debate in civil society around AV ethics, which in turn informs policy-making⁶⁷. More broadly, sci-fi-sci studies in this context have informed specific policy recommendations^{26,85} and subsequent legislative measures⁶⁸.

In particular, we used a conjoint design that enabled detecting the trade-offs participants were willing to make for these techno-ethical choices, a method that may be particularly useful to measure attitudes and behaviours driven by speculative technologies.

In hindsight, this was sci-fi-sci research, in the sense that we asked our participants to imagine a future technology, assigned them to various experimental treatments corresponding to hypothetical regulations of that technology, and recorded behavioural measures such as their support for the government that enacted the regulation, or their intention to purchase the regulated technology. These articles and many other subsequent behavioural articles about AVs allowed policy debates to take place ahead of technology developments. Fully autonomous AVs are still not available for purchase, but thanks to sci-fi-sci, we have made much progress on their ethical regulation^{26–28} and we are less likely to be blindsided as a society by their behavioural implications²⁹. In a similar vein, research on service robots has a long history of attempting

to predict the behavioural interactions people may have with futuristic robots, as well as the social implications of their introduction^{30,31} (see Box 3 for further details).

Before we consider other sci-fi-sci examples, we should first comment on the challenge of conducting a systematic literature review, or assessing the impact of this approach on technology development and regulation. Sci-fi-sci, unlike futures studies (a well-established field with dedicated journals, recognized methods and searchable keywords) lacks easy retrievability. It is not easily searchable, nor centralized in specific journals or conferences. Moreover, evaluating the influence of sci-fi-sci on technological development and regulation can be problematic, because tech companies (and sometimes policymakers) lack transparency about how they incorporate behavioural research in their decision-making. This problem is compounded by the fact that experimental research on the impact of speculative technologies may be increasingly shaped or conducted by tech companies with few incentives to make their findings public³². Given the applied importance of predicting the behavioural and social effects of emerging technologies, especially at a time when advances in artificial intelligence (AI) spawn both hope and fear³³, we need to understand the relative reluctance of behavioural scientists to engage in the experimental exploration of technology-driven futures. In other words, why has sci-fi-sci not yet become a well-established field?

Challenges for science fiction science

Behavioural scientists who seek to inform policy-making and technological regulation need to tackle the challenge of the ecological validity of their experiments (that is, the likelihood that the findings they obtain in the laboratory can predict behaviour in the real world) as well as the temporal validity of their findings (the endurance of their ecological validity into a changing future)^{34–37}. These challenges are especially problematic for sci-fi-sci experiments. Sci-fi-sci experiments cannot have ecological validity in a strict sense, because there is no ‘real world’ that the studies seek to generalize to. The world they try to generalize to does not exist yet, and may in fact never exist. Indeed, the challenge of temporal validity is inverted for sci-fi-sci experiments. For traditional experiments, the challenge of temporal validity is that ecological validity decays over time, from the present moment onward. For sci-fi-sci experiment, the hope is that their ecological validity will increase over time, in the sense that their findings will one day reflect real-world behaviour once their target technology is deployed for real.

Threats to such prospective ecological validity can occur in three areas. First, participants from the present may fail to simulate the behaviour of actual users of the technology. We know, for example, that people can have a hard time predicting their future emotional states³⁸ even when they are trying to picture situations they know well. A sci-fi-sci experiment may pose an even greater challenge to participants if it requires them to imagine their cognitive and emotional reactions to an unfamiliar future technology in an unfamiliar future context. Second, the depiction of the technology used by experimenters might be substantially different from the actual version once it is developed. For example, 20 years ago, prospective studies of the acceptability of nanotechnologies asked people how comfortable they would be with the nano-augmentation of cognitive capacities⁸⁹. Given that nanotechnologies have not developed in that direction, we can now recognize that the prospective validity of these questions was low. Third, the social context of the experiment may differ substantially from the social context at the time the technology is developed. Changes in social context can affect people’s attitudes and behaviours towards technology. For example, the COVID-19 pandemic increased acceptance of care robots³⁹ and shifts in religiosity can affect attitudes towards assisted reproduction technologies⁴⁰. Unexpected changes in social context between the sci-fi-sci experiment and the actual deployment of the technology may thus threaten prospective validity.

Box 3

Case study B: cooperating with autonomous agents

Period of sci-fi-sci studies. 2000s to present day.

Sci-fi-sci question. How should an autonomous agent (such as a robot or software agent) prioritize its own interests? This question was posited by the science fiction author Isaac Asimov in his third ‘law of robotics’: “A robot must protect its own existence as long as such protection does not conflict with the First or Second Law”⁸⁶. In *What Matters to a Machine?*, AI pioneer Drew McDermott imagined a robot that was tempted to break an ethical rule to further its owner’s interests⁸⁷.

Technological plausibility. Early software agents and robots were mere tools for performing specific tasks. They did not exhibit autonomy, nor did they face situations in which their goals conflicted with humans. By the early 2000s, we had various experimental demonstrations of autonomous social robots^{88–90}, and software agents that managed business processes⁹¹ (TRL=5–7). Scenarios requiring such agents to negotiate their own interests with humans became increasingly plausible.

Temporal proximity. Approximately two decades later, software agents became mainstream⁹² (TRL=9–10), and corporate investment in humanoid robotics reached US\$2.43 billion in 2023⁹³ (TRL=7–8). This led AI experts to demand that autonomous machines must learn to find common ground and cooperate with each other and with humans⁹⁴.

Magnitude of effect. The long-term socioeconomic effects of autonomous agents are difficult to predict⁹⁵. However, it is feasible to study the early dynamics of communication and cooperation between humans and autonomous agents when their interests are not fully aligned.

Sample of sci-fi-sci studies. About a decade ago, a series of studies explored whether autonomous agents, tasked with maximizing their own interest, can establish stable cooperation with humans⁹⁶. These experiments adapted paradigms from behavioural economics, such as the repeated prisoner’s dilemma. These studies revealed that humans were less likely to cooperate with benevolent machine agents than with humans⁹⁷, because humans consider it acceptable to exploit them⁹⁸. These tendencies are influenced by visual features of autonomous agents⁹⁹ as well as signals about emerging human-machine norms¹⁰⁰. Related work explored situations in which machine agents can exert authority over humans—for example, as managers¹⁰¹—and identified moral hazards that may arise because people exhibit obedience to the authority of a robot boss⁵³.

Subsequent relevance. The studies cited above may have seemed highly speculative just a few years ago. However, this has changed in the past two years with the sudden and rapid rise of conversational agents like ChatGPT¹⁰² and their numerous autonomous (agentic) implementations¹⁰³. Suddenly, we live in a world in which we interact frequently with autonomous customer service agents acting on behalf of other organizations¹⁰⁴. The challenge of establishing cooperation between humans and autonomous agents is already pervasive, but fortunately we had a head start.

These are formidable challenges. Without a set of accepted principles and methods to address them, behavioural scientists may have been dissuaded from producing what would otherwise have been useful research. In fact, it is plausible that the absence of established

approaches has created a self-perpetuating cycle that prevents behavioural scientists from conducting experiments on the effects of future technologies. Concerns about methodological issues may have made researchers hesitant to predict these effects—and this hesitation, in turn, may have prevented the formation of a diverse community of behavioural scientists dedicated to developing the principles and methods needed to study future technologies.

Nevertheless, researchers from several fields have experimented with methods that may minimize threats to prospective validity (Fig. 1). For instance, many studies have devised methods to simulate the experience of a technology in order to bring participants psychologically closer to the future that the researchers are attempting to study. Often, this type of simulation has been limited to vignettes: the technology is described to participants, leaving to their imagination how it would feel to use it. This has sometimes been the only feasible option—for example, for situations such as neuro-enhancement drugs—where the technology cannot be simulated otherwise^{41–44}. However, the more that contextual factors are left to the imagination of the participants, the more participants will fill in the gaps themselves, leading to unwanted variation. For example, some participants may project into the experiment the context of some science fiction works that they are familiar with⁴⁵. To limit such unwanted variations, experimenters need to provide as much contextual information as they can. One way to move beyond vignettes and to improve the consistency and verisimilitude of the simulated experience is to use of mock versions of futuristic apps^{46–48}. Although these mock-ups cannot execute the full range of capabilities of the technology that they are simulating, they embed an anticipated future technology within a context that participants are accustomed to. For example, Longoni and Cian⁴⁷ presented participants with a mobile app that purported to provide recommendations of an AI master chocolatier. Simulated apps may also enable the use of immersive behavioural measures, allowing participants to ‘stay in character’ as their future selves, rather than risking a jarring transition to standard survey questions that could undermine the effect of the simulation.

Researchers are also increasingly turning to virtual and augmented reality to further increase immersion. For example, the moral dilemmas of AVs were initially studied through vignettes with static images²³, then later with interactive computer graphics²⁵, and most recently with virtual reality^{49–51} in order to better situate participants within the expected reality of a world with pervasive driverless cars carrying out complex moral decisions. Other studies have used digital and even physical avatars to simulate the interaction with hypothetical AI-powered systems, advanced beyond the capabilities of current AI. For example, researchers have used ‘Wizard of Oz’ techniques, in which human confederates impersonated AI chatbots with responsive abilities that were at the time impossible for the chatbots to generate themselves⁵². This technique has also been used to give research participants the experience of interacting with speculative humanoid robots—for example, robot bosses whose blinking and neck movements are autonomously generated, but whose decisions and responses are controlled in real time by experimenters monitoring the interaction via cameras⁵³. Finally, at the far end of the fidelity spectrum, some research has made use of extremely elaborate staged settings that are inaccessible to most researchers—for example, the Mars missions simulations^{54–56} involved purpose-built habitats to mimic the confinement, isolation and resource scarcity of a future micro-society in a new environment.

These studies showcase some methodological solutions that researchers have found to the problem of running experiments on the future, and it remains to be seen how successful these attempts can be. In addition, although the core of sci-fi-sci revolves around experimental studies involving human participants, we recognize the potentially valuable role of computational methods such as Monte Carlo simulations and agent-based modelling in enriching our understanding of the future. These methods enable researchers to model complex systems and explore how assumptions about human behaviour and

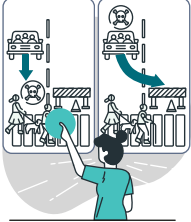


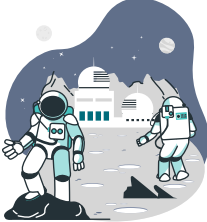
Approach	Text vignette	Multi-modal description	Virtual and augmented reality	Situational physical simulation	Environmental physical simulation
How it works	Via a text description, invite the participant to imagine a future scenario	Provide a visual aid to help the participant visualize the future scenario	Use virtual reality or augmented reality simulation of the future scenario	Immerse participants in a realistic physical simulation of a specific future situation	Immerse participants in a full physical simulation of a future environment
Concrete example of the simulation technique	<p>Here are some potential risks of nanotechnology. Which item is it most important to avoid?</p> <ol style="list-style-type: none"> 1. Losing your personal privacy to tiny new surveillance devices 2. Uncontrollable spread of self-replicating nano-sized robots. 3. ... 	<p>The Moral Machine: what should the self-driving car do?</p> 	<p>Virtual reality simulation of autonomous vehicle moral dilemmas</p> 	<p>Obedience to a robotic boss</p> 	<p>Mars Desert Research Station</p> 
Increasing fidelity →					

Fig. 1 | Examples of methods for simulating future scenarios. From left to right, approaches for simulating future scenarios are shown, from the lowest fidelity text vignettes, to the highest fidelity full environmental physical simulation.

technological adoption might play out over time. In the same spirit, AI simulations using autonomous programs (bots) programmed with different behavioural assumptions may sometimes serve as proxies for human participants, especially when exploring large-scale social dynamics that are impractical to study in laboratory settings. Integrating these computational approaches can complement the experimental data obtained from human participants, offering a richer view of potential futures shaped by speculative technologies.

As the sci-fi-sci literature grows, the community will trade perspectives and accumulate data on which methods work best for which purposes, and when. When is complex world-building necessary? When do simple vignettes suffice and when are laborious virtual or physical simulations beneficial? Conversely, when do these simulations risk creating demand effects linked to the experimenters’ own vision of the future? When does prospective validity benefit from having immersive measures in addition to immersive stimuli? The solution space also has considerable room to grow. As noted, these methods are scattered across various disciplines, from psychology to experiential ethnographic futures^{57,58} that rarely find themselves in conversation. Encouraging discussion among the diverse researchers who are keen to test behavioural hypotheses about future technologies could lead to the fertile recombination of disciplinary expertise, significantly increasing the methodological toolkit of this new community.

Still, most of these solutions will be focused on the first threat to prospective validity—that is, the challenge of eliciting responses from participants that accurately reflect how people would react in an anticipated future. They do not address the second and third challenges of the difficulty in accurately anticipating a technology and the social context in which it will be deployed. Here researchers need to balance the benefits that their research could bring with the uncertainty inherent in a speculative, prospective science. Nevertheless, certain topics will be laden with more uncertainty than others. Threats to prospective validity should thus be minimized by recognizing the features that can make topics less speculative and more amenable to study. In the next section, we consider some first guidelines for choosing such topics.

Topics for science fiction science

In her science fiction novel *Too Like the Lightning*, Ada Palmer⁵⁹ describes a future in which the hyper-mobility afforded by free, flying, supersonic

self-driving cars led to the collapse of nation states, which were replaced by global communities of like-minded individuals. Although this technology makes for a fantastic story, it would arguably be inadequate as the focus of a sci-fi-sci experiment. First, the development of such technology is very far from our current capabilities. This makes it difficult to imagine its plausible specifications, and even more difficult to envision the societal landscape by the time it might be invented. Second, geopolitical effects like the collapse of nation states unfold over a long period of time (years or even generations). Such effects cannot be properly measured within the brief duration of an experiment. Finally, the target technology is so disruptive that it transforms everything. This makes it difficult for participants to simulate the behaviour of future people if they not only have to picture a new technology, but also deal with its all-encompassing ramifications.

The above example illustrates how uncertainty around the future can become simply too great to conduct a valid experiment. This concept is often captured visually via a ‘futures cone’⁶⁰, which we have adapted for our purposes (Fig. 2). The cone of uncertainty represents the idea that as we project further into the future (or the past), the range of possible outcomes (or histories) broadens due to the increasing number of uncertain or unpredictable variables. This is true in fields that study past human behaviour (such as cognitive archaeology⁶¹, historical psychology⁶⁰ and economic history⁶²) as well as future behaviour (such as sci-fi-sci and future studies). We can adapt the futures cone to explore what makes an ideal topic for sci-fi-sci.

To start with, the further away a technology is in the future, the greater the cone of uncertainty. As a result, it is good practice to stick to the near future when designing a sci-fi-sci experiment. What constitutes the near future is of course up for debate, but one place to start is by studying the near-term behavioural effects of applications that seem to be almost within reach. For example, recent advances in large language models may enable widespread use of automated lie detection in interpersonal communication. Behavioural experiments can explore how such capabilities may disrupt existing social dynamics—for example, making people less inhibited in accusing others of lying⁶³. When it comes to technologies that do not yet exist, one heuristic is to consider the technological readiness level (TRL) of the target technology⁶⁴. The TRL scale, developed by NASA and adopted by many other agencies, assesses the maturity of a technology. It ranges from level 1, where only basic scientific principles are observed and potential applications are

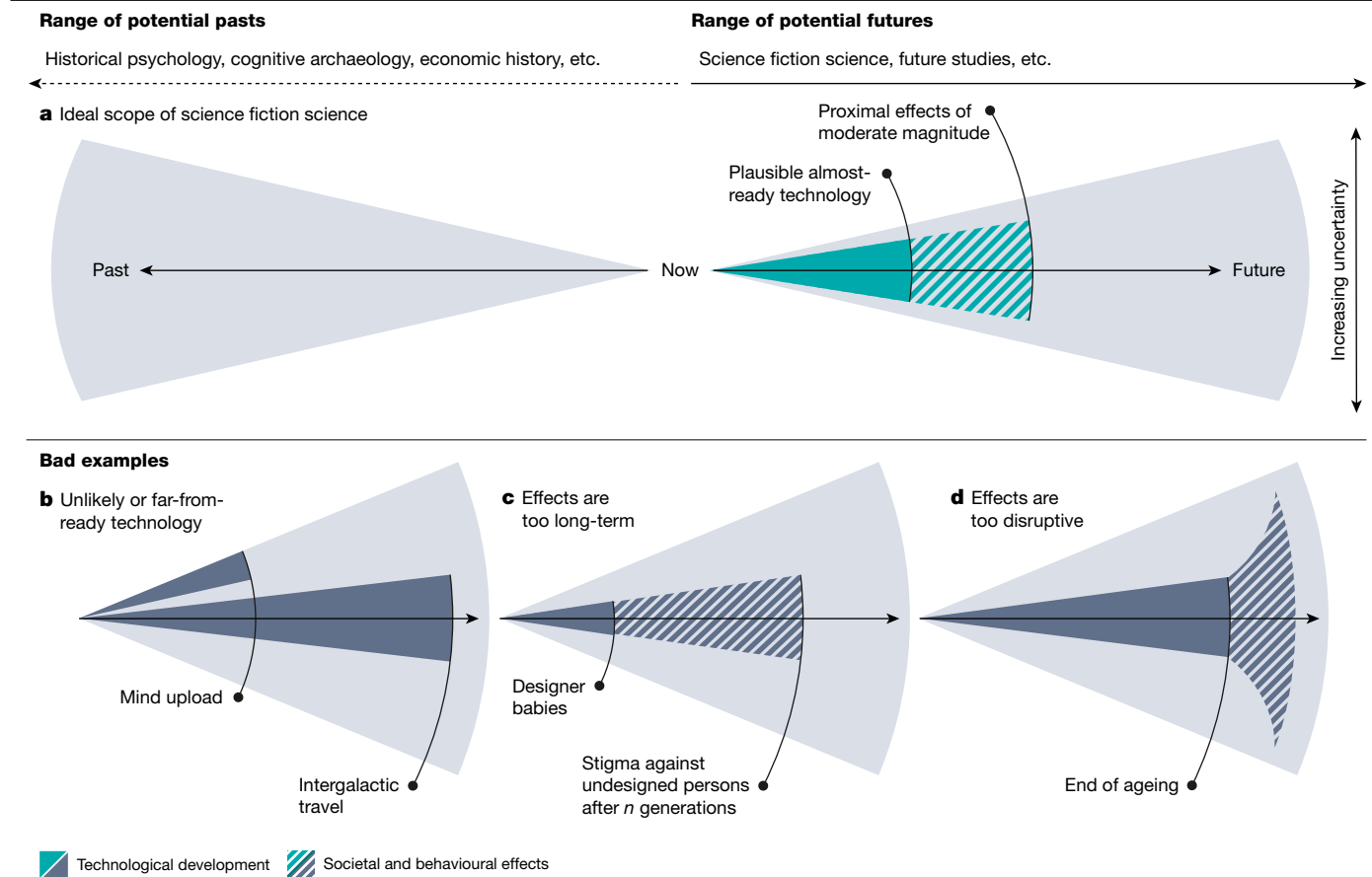


Fig. 2 | A modified futures cone for sci-fi-sci. **a**, The ideal domain for sci-fi-sci is to study proximal, moderate magnitude effects of plausible, almost-ready technologies. **b–d**, Deviations from the ideal depicted in **a**. Topics less amenable to sci-fi-sci include unlikely or far from ready technologies (**b**); societal and

behavioural effects that only materialize too far into the future after the introduction of the technology (**c**); and technologies whose effects are so disruptive that they would change too many aspects of society all at once (**d**).

not well formulated, to level 9, where the technology is validated in its intended operational environment. Although sci-fi-sci studies may not need to wait until a technology reaches level 9, it seems reasonable to focus on technologies that are at least 4 (small-scale prototype) on this scale. The further along a technology is on the TRL scale, the more likely its sci-fi-sci version is to resemble its eventual real version.

In addition, when running a sci-fi-sci experiment, we must consider not only when a technology will appear but also the timescale of the effects that we want to investigate. The longer it takes for these effects to unfold, the larger the cone of uncertainty around them, and the greater the threat to validity. Therefore, it is probably good practice to study the proximal effects of the technology rather than those that unfold over a timescale beyond that of a behavioural experiment. For example, we could study the social acceptability of future parents selecting or deselecting various traits when choosing an embryo for implantation. However, it would be much more difficult to study how their choices might affect the social stigma attached to the deselected traits over time.

Finally, there are some technological breakthroughs (for example, anti-ageing treatments that could push human longevity into centuries) that would be so disruptive that their appearance may change everything, exponentially increasing the cone of uncertainty around them. Whereas science fiction writers may naturally gravitate towards these game-changing technologies because of their narrative potential, sci-fi-sci experimenters may prefer to focus on technologies that bring changes of moderate magnitude. Although there is no analogue to the TRL for quantifying the predicted impact of a future technology, Coccia's 'scale of innovative intensity'⁶⁵ categorizes historical technologies

into three tiers of impact: low, medium and high. The high-impact category is further divided into 'very strong' (for example, the Internet) and 'revolutionary' (for example, steam power or electricity) technologies with such profound effects that they reshape nearly every sector of the economy and touch nearly every individual on the planet. Again, as a rough heuristic, sci-fi-sci may be best targeted at emerging technologies that would fit in the low- and medium-impact tiers of the Coccia scale. Recent examples of technological innovations in this range include personalized algorithms for media content, direct-to-consumer genetic testing, SMS texting, credit cards and other forms of digital cashless payment, 24-hour news channels and in vitro fertilization. Each of these was an incremental technological advance that could have been predicted in the years preceding its emergence, but each also had pronounced and often unintended societal consequences.

This is not to say that sci-fi-sci should fully avoid sweepingly transformative technologies such as AI, which will almost certainly have very strong and possibly revolutionary impact. Sci-fi-sci can and should engage with AI, but it is challenging to design experiments around the technology's most ambitious, civilization-changing scenarios. Instead, sci-fi-sci researchers are better off breaking down AI into more manageable components and narrower applications, such as self-driving technology or autonomous medical triage—just as researchers in the near past would have found it more manageable to design experiments around the impact of a transformative but evolutionary technology such as social media rather than a paradigm-shifting technology like the Internet (or electricity). Such scenarios remain close enough to current conditions that researchers can realistically simulate them, vary key parameters and measure meaningful responses without collapsing

under the weight of unbounded speculation. In this sense, we do not propose sidestepping AI altogether, but rather targeting more delimited forms of AI-driven technologies whose behavioural impacts can be credibly tested in a laboratory context.

Although it is difficult to define formal guidelines for topic selection (in the broad sense of choosing a technology of interest and defining the exact scenarios for investigation), it is an area where synergy between sci-fi-sci and futures studies will be especially important. Indeed, futures studies researchers have already established a trove of ready topics through methods such as back casting, whereby futurologists trace back the technological path (and societal response) that would lead to a desirable outcome. Their topics have been qualitatively derived and now stand ready for quantitative testing via the sci-fi-sci method. Moreover, in order to generate realistic and comprehensive scenarios grounded in domain-specific knowledge, sci-fi-sci researchers will find great value in collaborating with governmental and non-governmental organizations that engage in scenario planning; with experts of methods such as prospective hindsight and pre-mortem studies; and with practitioners who help ensure that experimental designs capture critical variables and potential harms that might otherwise be overlooked. By leveraging the detailed narratives and insights from futures studies, sci-fi-sci can create more meaningful and effective experiments, enhancing its ability to anticipate and study the impacts of speculative technologies.

Final remarks

Science fiction writers give us a glimpse of possible futures by telling the stories of how humans might be changed by new technologies. Sci-fi-sci applies the scientific method to the science fiction project by immersing research participants into controlled variations of the future and collecting quantitative data on their attitudes and behavioural responses. The seeds of sci-fi-sci were planted before this article: researchers from various fields are already providing immersive simulations of the future or asking research participants to express preferences and opinions about the technologies of the future. Accordingly, we do not claim to have invented a new field, but we hope that researchers in disparate fields who are already studying the behaviour of future humans will have an easier time finding each other with a flag to rally around.

We also hope that new research communities will discover the sci-fi-sci project and bring their expertise to this multidisciplinary enterprise. Studying the behaviour of future humans interacting with future technology in a future social world raises unusual challenges for behavioural scientists, which call for unconventional methods. Institutional mechanisms that can help overcome these challenges include higher risk tolerance among research funding organizations to create programmes that are explicitly designed for sci-fi-sci projects despite the substantial uncertainties involved; community-building meetings to facilitate the sharing of experiences and best practices; and coordinated efforts to establish rigorous standards for assessing sci-fi-sci experimental methods and findings. In particular, sci-fi-scientists will need to devise best practices to communicate their quantitative findings in order to avoid conveying a false sense of precision to stakeholders who may overfixate on numerical estimates that come with large uncertainty.

We hope that our review of these methods, their rationales and their limitations will encourage sci-fi scientists to make bold methodological choices and a convincing case for these bold choices. The future of sci-fi-sci will depend on the quality of its methods.

1. Braghieri, L., Levy, R. & Makarin, A. Social media and mental health. *Am. Econ. Rev.* **112**, 3660–3693 (2022).
2. Van Bavel, J. J., Robertson, C. E., Del Rosario, K., Rasmussen, J. & Rathje, S. Social media and morality. *Annu. Rev. Psychol.* **75**, 311–340 (2024).

3. Bail, C. Social-media reform is flying blind. *Nature* **603**, 766 (2022).
4. Frewer, L. et al. Societal aspects of genetically modified foods. *Food Chem. Toxicol.* **42**, 1181–1193 (2004).
5. Siegrist, M. & Hartmann, C. Consumer acceptance of novel food technologies. *Nat. Food* **1**, 343–350 (2020).
6. Scott, S. E., Inbar, Y. & Rozin, P. Evidence for absolute moral opposition to genetically modified food in the United States. *Perspect. Psychol. Sci.* **11**, 315–324 (2016).
7. Fernbach, P. M., Light, N., Scott, S. E., Inbar, Y. & Rozin, P. Extreme opponents of genetically modified foods know the least but think they know the most. *Nat. Hum. Behav.* **3**, 251–256 (2019).
8. Cobb, M. D. & Macoubrie, J. Public perceptions about nanotechnology: risks, benefits and trust. *J. Nanopart. Res.* **6**, 395–405 (2004).
9. Lee, C.-J., Scheufele, D. A. & Lewenstein, B. V. Public attitudes toward emerging technologies: examining the interactive effects of cognitions and affect on public attitudes toward nanotechnology. *Sci. Commun.* **27**, 240–267 (2005).
10. Danaher, J. & Sætra, H. S. Mechanisms of techno-moral change: a taxonomy and overview. *Ethical Theory Moral Pract.* **26**, 763–784 (2023).
This article offers a precise and generative taxonomy of how technology reshapes moral life, providing a conceptual foundation for designing sci-fi-sci scenarios with mechanistic clarity.
11. Goldin, C. & Katz, L. F. The power of the pill: oral contraceptives and women's career and marriage decisions. *J. Pol. Econ.* **110**, 730–770 (2002).
12. Bailey, M. J. More power to the pill: the impact of contraceptive freedom on women's life cycle labor supply. *Q. J. Econ.* **121**, 289–320 (2006).
13. Baker, R. *Before Bioethics* (Oxford Univ. Press, 2013).
14. Shariff, A., Green, J. & Jettinghoff, W. The privacy mismatch: evolved intuitions in a digital world. *Curr. Dir. Psychol. Sci.* **30**, 159–166 (2021).
15. Zuboff, S. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (PublicAffairs, 2019).
16. Collingridge, D. *The Social Control of Technology* (Pinter, 1980).
This foundational work diagnoses the dilemma of control in technological development, which sci-fi-sci attempts to tackle by generating early empirical insights before lock-in occurs.
17. Fergnani, A. Mapping futures studies scholarship from 1968 to present: a bibliometric review of thematic clusters, research trends, and research gaps. *Futures* **105**, 104–123 (2019).
18. Kuosa, T. Evolution of futures studies. *Futures* **43**, 327–336 (2011).
19. Danaher, J. & Sætra, H. S. Technology and moral change: the transformation of truth and trust. *Ethics Inf. Technol.* **24**, 35 (2022).
20. Hopster, J. K. & Maas, M. M. The technology triad: disruptive AI, regulatory gaps and value change. *AI Ethics* **4**, 1051–1069 (2024).
21. Brey, P. *Ethics of Emerging Technology* 175–191 (Rowman & Littlefield, 2017).
22. Pohl, F. The great new inventions. *Galaxy* **27**, 6 (1968).
23. Bonnefon, J.-F., Shariff, A. & Rahwan, I. The social dilemma of autonomous vehicles. *Science* **352**, 1573–1576 (2016).
24. Shariff, A., Bonnefon, J.-F. & Rahwan, I. Psychological roadblocks to the adoption of self-driving vehicles. *Nat. Hum. Behav.* **1**, 694–696 (2017).
25. Awad, E. et al. The Moral Machine experiment. *Nature* **563**, 59–64 (2018).
This study is a classic example of sci-fi-sci, experimentally probing public moral preferences for a speculative technology (fully autonomous vehicles) through a massive global dataset of 40 million decisions.
26. Bonnefon, J.-F. et al. *Ethics of connected and automated vehicles: recommendations on road safety, privacy, fairness, explainability and responsibility* (European Commission, 2020).
27. Luetge, C. The German ethics code for automated and connected driving. *Philos. Technol.* **30**, 547–558 (2017).
28. Santoni de Sio, F. The European Commission report on ethics of connected and automated vehicles and the future of ethics of transportation. *Ethics Inf. Technol.* **23**, 713–726 (2021).
29. Adnan, N. Exploring the future: a meta-analysis of autonomous vehicle adoption and its impact on urban life and the healthcare sector. *Transp. Res. Interdiscip. Persp.* **26**, 101110 (2024).
30. Tussyadiah, I. A review of research into automation in tourism: launching the annals of tourism research curated collection on artificial intelligence and robotics in tourism. *Ann. Tour. Res.* **81**, 102883 (2020).
31. Zeng, Y., Liu, X., Zhang, X. & Li, Z. Retrospective of interdisciplinary research on robot services (1954–2023): from parasitism to symbiosis. *Technol. Soc.* **78**, 102636 (2024).
32. Benkler, Y. Don't let industry write the rules for AI. *Nature* **569**, 161–162 (2019).
33. Cave, S. & Dihal, K. Hopes and fears for intelligent machines in fiction and reality. *Nat. Mach. Intell.* **1**, 74–78 (2019).
34. Lazer, D. et al. Computational social science: obstacles and opportunities. *Science* **369**, 1060–1062 (2020).
35. Lazer, D. et al. Meaningful measures of human society in the twenty-first century. *Nature* **595**, 189–196 (2021).
36. Munger, K. The limited value of non-replicable field experiments in contexts with low temporal validity. *Soc. Media Soc.* **5**, 2056305119859294 (2019).
37. Munger, K. Temporal validity as meta-science. *Res. Politics* **10**, 20531680231187271 (2023).
This article unpacks the concept of temporal validity, an essential concern for sci-fi-sci, as it exposes the limits of applying present-day empirical knowledge to future contexts.
38. Wilson, T. D. & Gilbert, D. T. Affective forecasting. *Adv. Exp. Soc. Psychol.* **35**, 345–411 (2003).
39. Schönmann, M., Bodenschatz, A., Uhl, M. & Walkowitz, G. Contagious humans: a pandemic's positive effect on attitudes towards care robots. *Technol. Soc.* **76**, 102464 (2024).
40. Inhorn, M. C. & Birenbaum-Carmeli, D. Assisted reproductive technologies and culture change. *Annu. Rev. Anthropol.* **37**, 177–196 (2008).
41. Dinh, C. T., Humphries, S. & Chatterjee, A. Public opinion on cognitive enhancement varies across different situations. *Am. J. Bioethics* **11**, 224–237 (2020).

42. Mihailov, E., López, B. R., Cova, F. & Hannikainen, I. R. How pills undermine skills: moralization of cognitive enhancement and causal selection. *Conscious. Cogn.* **91**, 103120 (2021).
43. Sattler, S. et al. Neuroenhancements in the military: a mixed-method pilot study on attitudes of staff officers to ethics and rules. *Neuroethics* **15**, 11 (2022).
44. Lucas, S., Douglas, T. & Faber, N. S. How moral bioenhancement affects perceived praiseworthiness. *Bioethics* **38**, 129–137 (2024).
45. Laakasuo, M. et al. What makes people approve or condemn mind upload technology? untangling the effects of sexual disgust, purity and science fiction familiarity. *Palgrave Commun.* **4**, 84 (2018).
46. Salganik, M. J., Dodds, P. S. & Watts, D. J. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**, 854–856 (2006).
47. Longoni, C. & Cian, L. Artificial intelligence in utilitarian vs. hedonic contexts: the ‘word-of-machine’ effect. *J. Mark.* **86**, 91–108 (2022).
48. Köbis, N. et al. Artificial intelligence can facilitate selfish decisions by altering the appearance of interaction partners. Preprint at <https://doi.org/10.48550/arXiv.2306.04484> (2023).
49. Benvegñù, G., Pluchino, P. & Garnberini, L. Virtual morality: using virtual reality to study moral behavior in extreme accident situations. In *2021 IEEE Virtual Reality and 3D User Interfaces* 316–325 (IEEE, 2021).
50. Sütfeld, L. R., Ehinger, B. V., König, P. & Pipa, G. How does the method change what we measure? comparing virtual reality and text-based surveys for the assessment of moral decisions in traffic dilemmas. *PLoS ONE* **14**, e0223108 (2019).
51. Faulhaber, A. K. et al. Human decisions in moral dilemmas are largely described by utilitarianism: Virtual car driving study provides guidelines for autonomous driving vehicles. *Sci. Eng. Ethics* **25**, 399–418 (2019).
52. Riek, L. D. Wizard of Oz studies in HRI: a systematic review and new reporting guidelines. *J. Hum. Robot Interact.* **1**, 119–136 (2012).
53. Aroyo, A. M. et al. Will people morally crack under the authority of a famous wicked robot? In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication* 35–42 (IEEE, 2018).
54. Bishop, S. L., Kobrick, R., Battler, M. & Binsted, K. Fmars 2007: stress and coping in an Arctic Mars simulation. *Acta Astronautica* **66**, 1353–1367 (2010).
55. Alfano, C. A., Bower, J. L., Cowie, J., Lau, S. & Simpson, R. J. Long-duration space exploration and emotional health: recommendations for conceptualizing and evaluating risk. *Acta Astronautica* **142**, 289–299 (2018).
- This article reviews methods for forecasting emotional health risks during a Mars mission, an extreme case of sci-fi sci aimed at predicting human responses in radically novel, high-stakes environments.**
56. Riva, P., Rusconi, P., Pancani, L. & Chterev, K. Social isolation in space: an investigation of LUNARK, the first human mission in an Arctic Moon analog habitat. *Acta Astronautica* **195**, 215–225 (2022).
57. Candy, S. & Potter, C. Design and Futures, vol. 1. *J. Futures Stud.* (2019).
58. Candy, S. & Potter, C. Design and Futures, vol. 2. *J. Futures Stud.* (2019).
- The two volumes of this special issue are a treasure trove of innovative methods for designing immersive future simulations, offering sci-fi-sci researchers a rich toolbox for experimental protocols grounded in a tangible experience.**
59. Palmer, A. *Too Like the Lightning* (Tor Books, 2016).
60. Gall, T., Vallet, F. & Yannou, B. How to visualise futures studies concepts: revision of the futures cone. *Futures* **143**, 103024 (2022).
61. Coolidge, F. L., Wynn, T., Overmann, K. A. & Hicks, J. M. in *Human Paleoneurology* (ed. Bruner, E.) 177–208 (Springer, 2015).
62. North, D. C. Structure and performance: the task of economic history. *J. Econ. Lit.* **16**, 963–978 (1978).
63. von Schenk, A., Klockmann, V., Bonnefon, J.-F., Rahwan, I. & Köbis, N. Lie detection algorithms disrupt the social dynamics of accusation behavior. *iScience* **27**, 110201 (2024).
64. Héder, M. From NASA to EU: the evolution of the TRL scale in public sector innovation. *Innov. J.* **22**, 1–23 (2017).
- This article unpacks the evolution and potential misuses of the technology readiness level scale, which can help sci-fi-sci researchers to choose speculative technologies that are still uncertain, but not untethered from reality.**
65. Coccia, M. Measuring intensity of technological change: the seismic approach. *Technol. Forecast. Soc. Change* **72**, 117–144 (2005).
66. Goodall, N. J. Ethical decision making during automated vehicle crashes. *Transp. Res. Rec.* **2424**, 58–65 (2014).
67. Hess, D. J. Incumbent-led transitions and civil society: autonomous vehicle policy and consumer organizations in the united states. *Technol. Forecast. Soc. Change* **151**, 119825 (2020).
68. Kriebitz, A., Max, R. & Lütge, C. The German act on autonomous driving: why ethics still matters. *Philos. Technol.* **35**, 29 (2022).
69. Mac Sithigh, D. & Siemens, M. The Chinese social credit system: a model for other countries? *Mod. L. Rev.* **82**, 1034–1071 (2019).
70. Orgad, L. & Reijers, W. A *Dystopian Future? The Rise of Social Credit Systems*. Technical Report RSCAS 2019/94 (Robert Schuman Centre for Advanced Studies, 2019).
71. Tirole, J. Digital dystopia. *Am. Econ. Rev.* **111**, 2007–2048 (2021).
72. Purcell, Z. A. & Bonnefon, J.-F. Humans feel too special for machines to score their morals. *PNAS Nexus* **2**, pgad179 (2023).
73. *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts* <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206> (European Commission, 2021).
74. Turley, P. et al. Problems with using polygenic scores to select embryos. *N. Engl. J. Med.* **385**, 78–86 (2021).
75. Anomaly, J. *Creating Future People: The Ethics of Genetic Enhancement* (Taylor & Francis, 2020).
76. Anomaly, J., Gyngell, C. & Savulescu, J. Great minds think different: preserving cognitive diversity in an age of gene editing. *Bioethics* **34**, 81–89 (2020).
77. Gyngell, C. & Douglas, T. Stocking the genetic supermarket: reproductive genetic technologies and collective action problems. *Bioethics* **29**, 241–250 (2015).
78. Cavaliere, G. Ectogenesis and gender-based oppression: resisting the ideal of assimilation. *Bioethics* **34**, 727–734 (2020).
79. Hooton, V. & Romanis, E. C. Artificial womb technology, pregnancy, and EU employment rights. *J. L. Biosci.* **9**, lsac009 (2022).
80. Horn, C. Ectogenesis, inequality, and coercion: a reproductive justice-informed analysis of the impact of artificial wombs. *BioSocieties* **18**, 523–544 (2023).
81. MacKay, K. The ‘tyranny of reproduction’: could ectogenesis further women’s liberation? *Bioethics* **34**, 346–353 (2020).
82. Lin, P. in *Autonomous Driving: Technical, Legal and Social Aspects* (eds Maurer, M.) 69–85 (Springer Vieweg, 2016).
83. Milakis, D. Long-term implications of automated vehicles: an introduction. *Transp. Rev.* **39**, 1–8 (2019).
84. Soteropoulos, A., Berger, M. & Ciari, F. Impacts of automated vehicles on travel behaviour and land use: an international review of modelling studies. *Transp. Rev.* **39**, 29–49 (2019).
85. Fernández Llorca, D. & Gómez, E. *Trustworthy Autonomous Vehicles—Assessment Criteria for Trustworthy AI in the Autonomous Driving Domain* (European Union, 2021).
86. Asimov, I. *Robot* (Gnome Press, 1950).
87. McDermott, D. in *Machine Ethics* (eds Anderson, M. & Anderson, S. L.) 88–114 (Cambridge Univ. Press, 2011).
88. Hirai, K., Hirose, M., Haikawa, Y. & Takenaka, T. The Development of Honda Humanoid Robot. In *Proc. 1998 IEEE International Conference on Robotics and Automation* Vol. 2, 1321–1326 (IEEE, 1998).
89. Ishiguro, H. et al. Robovie: an interactive humanoid robot. *Ind. Robot* **28**, 498–504 (2001).
90. Breazeal, C. *Designing Sociable Robots* (MIT Press, 2004).
91. Jennings, N. R., Norman, T. J., Faratin, P., O’Brien, P. & Odgers, B. Autonomous agents for business process management. *Appl. Artif. Intel.* **14**, 145–189 (2000).
92. Reijers, H. A. Business process management: the evolution of a discipline. *Comput. Ind.* **126**, 103404 (2021).
93. *Humanoid Robot Market Size, Share & Industry Analysis, by Motion Type (Biped and Wheel Drive), by Component (Hardware and Software), by Application (Industrial, Household, and Services), and Regional Forecast 2024–2032*. Technical Report (Fortune Business Insights, 2024).
94. Dafoe, A. et al. Cooperative AI: machines must learn to find common ground. *Nature* **593**, 33–36 (2021).
95. Acemoglu, D. & Restrepo, P. Automation and new tasks: how technology displaces and reinstates labor. *J. Economic Perspect.* **33**, 3–30 (2019).
96. Crandall, J. W. et al. Cooperating with machines. *Nat. Commun.* **9**, 233 (2018).
97. Ishowo-Oloko, F. et al. Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nat. Mach. Intell.* **1**, 517–521 (2019).
98. Karpus, J., Krüger, A., Verba, J. T., Bahrami, B. & Deroy, O. Algorithm exploitation: humans are keen to exploit benevolent AI. *iScience* **24**, 102679 (2021).
99. Oudah, M., Makovi, K., Gray, K., Battu, B. & Rahwan, T. Perception of experience influences altruism and perception of agency influences trust in human–machine interactions. *Sci. Rep.* **14**, 12410 (2024).
100. Makovi, K., Sargsyan, A., Li, W., Bonnefon, J.-F. & Rahwan, T. Trust within human–machine collectives depends on the perceived consensus about cooperative norms. *Nat. Commun.* **14**, 3108 (2023).
101. Dong, M., Bonnefon, J.-F. & Rahwan, I. Toward human-centered AI management: methodological challenges and future directions. *Technovation* **131**, 102953 (2024).
102. Bubeck, S. et al. Sparks of artificial general intelligence: early experiments with GPT-4. Preprint at <https://doi.org/10.48550/arXiv.2303.12712> (2023).
103. Wu, Q. et al. Autogen: enabling next-gen LLM applications via multi-agent conversation framework. In *First Conference on Language Modeling* <https://openreview.net/forum?id=BAakY1hNKS> (OpenReview, 2024).
104. Bamberger, S., Clark, N., Ramachandran, S. & Sokolova, V. How generative AI is already transforming customer service. *Boston Consulting Group* www.bcg.com/publications/2023/how-generative-ai-transforms-customer-service (2023).

Acknowledgements J.-F.B. acknowledges support from grants ANR-19-PI3A-0004, ANR-17-EURE-0010 and ANR-22-CE26-0014-01, and the research foundation TSE–Partnership. A.S. acknowledges a Canada 150 Research Chair from the Social Science Research Council of Canada.

Competing interests The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Iyad Rahwan, Azim Shariff or Jean-François Bonnefon.

Peer review information *Nature* thanks Kristian Hammond and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© Springer Nature Limited 2025