"The Science Fiction Science Method"

Iyad Rahwan, Azim Shariff and Jean-François Bonnefon

Toulouse
School of
Economics

The Science Fiction Science Method

Iyad Rahwan[1,*], Azim Shariff[2,*], and Jean-François Bonnefon[3,*]

[1]Center for Humans & Machines, Max Planck Institute for Human Development, Berlin, Germany

[2]Department of Psychology, University of British Columbia, Vancouver, Canada

[3]Toulouse School of Economics, CNRS (TSM-R), University of Toulouse-Capitole, France

[*]Correspondence: `rahwan@mpib-berlin.mpg.de`, `shariff@psych.ubc.ca`,
`jean-francois.bonnefon@tse-fr.eu`

Abstract

Predicting the social and behavioral impact of future technologies, before they are achieved, would allow 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 which uses experimental methods to simulate future technologies, and collect quantitative measures of the attitudes and behaviors of participants assigned to controlled variations of the future. We call this method 'science fiction science'. We suggest that the reason why this method has not been fully embraced yet, despite its potential benefits, is that experimental scientists may be reluctant to engage in work facing such serious validity threats as science fiction science. To address these threats, we consider possible constraints on the kind of technology that science fiction science may study, as well as the unconventional, immersive methods that science fiction science may require. We seek to provide perspective on the reasons why this method has been marginalized for so long, what 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.

Main

Imagine that behavioral scientists managed to predict the effect of social media on mental health and democratic life, before social media actually existed. That is, imagine they 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 issues [1], or a focus on moral outrage, resulting in exaggerated feelings of polarization [2]. 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 effects [3].

Social media is just one example of a technology that could have benefited from early, speculative behavioral 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 failed to anticipate the psychological concerns of the public about this technology [4], which led to a persistent cycle of mistrust that has not been fully resolved even 30 years later [5]. Today, many people still manifest absolute moral opposition to genetically modified foods [6], although they understand very little about them [7]. It is of course impossible to claim that history would have been different if behavioral 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 behavioral impact of future technologies, it remains uncommon to do so with the experimental methods and quantitative measures typical of contemporary behavioral science. Consequently, we advocate for what we term *science fiction science:* the application of the scientific method to the science fiction project of anticipating behavioral and social changes driven by a speculative technology. After

providing a more detailed definition of science-fiction science, 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 science fiction science, 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 behavioral experiments centered 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 do [10]. 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 brain-dead patients changed the way doctors and patient think of the ethics of ending life [13]. Yet other technologies transformed society because they were *not* accompanied by significant changes in behavior and social norms. For example, sophisticated tracking technologies allowed tech giants to accumulate vast amount of personal data, leading to a commodification of human experience, because people failed to adjust their behaviors 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 behaviors, norms, and institutions now deeply entrenched. In contrast, we have much power to steer and regulate technologies which 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 (see box *Examples of policy debates about nascent or speculative technologies*). This tension

> ### Examples of policy debates about nascent or speculative technologies.
>
> **Connected and autonomous vehicles (CAVs)** 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 scenarios [66]. These debates have led to significant mobilization by Civil Society [67], and specific policy recommendations [26], some of which have informed legislation [68].
>
> **Social credit systems** Governments do not yet have the capacity to deploy AI systems that would monitor every behavior of all their citizens in real time, in order to calculate and release social scores; but the behavioral, social and political impact of these hypothetical systems is already the object of much speculation [69, 70, 71, 72], to the point where the European Union is considering a preemptive ban of this technology for its member governments [73].
>
> **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 offspring [74]. 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, 76, 77].
>
> **Ectogenesis** Medical research has not yet achieved ectogenesis (i.e., the artificial gestation of human foetuses), but the field of speculative bioethics is already debating whether this potential technology might end gender-based oppression, or exacerbate other inequalities, and to which extent it could disrupt social and legal norms around motherhood and employment rights [78, 79, 80, 81].

is known as the Collingridge control dilemma [16]: 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 panelists [17]. Future studies typically attempt to compose various scenarios describing the future, based on expert in-

tuition, and then work backwards to imagine which actions or events may lead to these scenarios [18]. 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 behavioral measures [21]. 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, behavioral counterpart to futures studies and other qualitative explorations of the impact of speculative technologies.

Science fiction science

Let us summarize our problem so far. We start from a technology that is not available yet, but is likely to become available in the future, with some ill-defined capabilities and limitations. We believe it may transform individual behaviors, social norms, and institutions, with a mixture of desirable and undesirable effects. The potential magnitude of these effects is large enough for us to worry about them even before the technology is achieved, in the hope that we can steer its development toward 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 behavioral data about what people will think and do when it becomes available, no controlled experiments, no tentative quantification of its positive and negative effects.

To collect such data, we need to develop methods and principles for a speculative behavioral science. We propose to call this method *science fiction science* (or sci-fi-sci for short), 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 behavioral 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

behavioral experiments: science fiction tells the stories of future humans, but sci-fi-sci attempts to study them in the lab.

More precisely, sci-fi-sci is the application of the scientific method to anticipate and study the behavioral, 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 behaviors. This approach employs rigorous experimental methods—including control and treatment conditions with independent and orthogonal variable manipulation, to not only survey attitudes but also measure direct behaviors 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 behavior. It serves as the quantitative, experimental, and behavioral 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 *Case Study A* box for further details). Between 2016 and 2018, we published several articles on fully autonomous vehicles, 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, 24, 25]. In particular, we used a conjoint design which allowed for detecting the trade-offs participants were willing to make for these techno-ethical choices, a method which may be particularly useful to measure attitudes and behaviors driven by speculative technologies.

Case Study A: Ethical dilemmas of autonomous vehicles (AVs)

Period of Sci-Fi-Sci studies: 2016–today

Sci-fi-sci Question: How do citizens and consumers wish AVs to prioritize the safety of different road users in unavoidable accidents? Technology ethicists [82] and transportation experts [66] were already discussing such possibilities before the mid-2010s.

Technological Plausibility: Many proof-of-concept tests took place through the DARPA Urban Challenge in 2007 (Technology Readiness Level=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 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 behavior or even the value of time [83]. Many of these outcomes, often based on computational simulation, are highly sensitive to modeling assumptions [84], and are thus difficult to study behaviorally. In 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 vignettes [23] highlighting the disconnect between the preferences of citizens (AVs should save as many lives as possible) and those of consumers (AVs should prioritize passengres). The Moral Machine experiment [25] used a conjoint design with visualizations of binary choices, crowdsourcing over 40 million decisions from people world wide, 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, 50, 51].

Subsequent relevance: By engaging millions of citizens world-wide, the Moral Machine experiment itself has contributed to a significant Civil Society debate around AV ethics, which in turn informs policy making [67]. More broadly, sci-fi-sci studies in this context have informed specific policy recommendations [26, 85], and subsequent legislations [68].

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 behavioral 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 behavioral articles about autonomous vehicles allowed policy debates to take place ahead of technology developments: Fully autonomous vehicles are still not available for purchase, but thanks to sci-fi-sci, we have made much progress on their ethical regulation [26, 27, 28], and we are less likely to be blindsided as a society by their behavioral implications [29]. In like vein, research on service robots has a long history of attempting to predict the behavioral interactions people may have with futuristic robots, as well as the social implications of their introduction [30, 31] (see *Case Study B* box 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 behavioral 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 public [32]. Given the applied importance of predicting the behavioral and social effects of emerging technologies, especially at a time when advances in Artificial Intelligence spawn both hope and fear [33], we need to understand the relative reluctance of behavioral 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?

Case Study B: Cooperating with autonomous agents

Period of Sci-Fi-Sci studies: 2000s–today

Sci-fi-sci Question: How should an autonomous agent (e.g. robot, software agent) prioritize its own interests? This question was posited by sci-fi author Isaac Asimov in his 3rd law: 'A robot must protect its own existence as long as such protection does not conflict with the First or Second Law' [86]. In a 2011 article titled 'What matters to a machine?', AI pioneer Drew McDermott imagined a robot tempted to break an ethical rule to further its owners interests [87].

Technological Plausibility: Early software agents and robots were mere tools 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, 89, 90], and software agents that manage business processes [91] (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 mainstream [92] (TRL=9–10), and corporate investment in humanoid robotics reached USD 2.43 billion in 2023 [93] (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 humans [94].

Magnitude of Effect: The long-term socioeconomic effects of autonomous agents are difficult to predict [95]. 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 humans [96]. These experiments adapted paradigms from behavioral economics, such as the Repeated Prisoner's Dilemma. These studies revealed that humans were less likely to cooperate with benevolent machine agents compared to humans [97], because humans consider it acceptable to exploit them [98]. These tendencies are influenced by visual features of autonomous agents [99] as well as signals about emerging human-machine norms [100]. Related work explored situations in which machine agents can exert authority over humans, e.g. as managers [101], and identified moral hazards that may arise because people exhibit obedience to the authority of a robot boss [53].
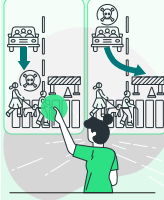
Subsequent relevance: The studies cited above may have seemed highly speculative just a few years ago. But this all changed in the last two years with the sudden and rapid rise of conversational agents like ChatGPT [102], and their numerous autonomous ('agentic') implementations [103]. Suddenly, we live in a world where we interact frequently with autonomous customer service agents acting on behalf of other organizations [104]. The challenge of establishing cooperation between humans and autonomous agents is already pervasive, but fortunately we had a head start.

Challenges for science fiction science

Behavioral scientists who seek to inform policymaking and technological regulation need to tackle the challenge of the ecological validity of their experiments (i.e., the likelihood that the findings they obtain in the lab can predict behavior in the real world) as well as the temporal validity of their findings (i.e., the endurance of their ecological validity into a changing future) [34, 35, 36, 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 behavior, 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 behavior of actual users of the technology. We know, for example, that people can have a hard time predicting their future emotional states [38], 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 significantly 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 capacities [8, 9]. 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 significantly from the social context at the time the technology is developed. Changes in social context can significantly impact people's attitudes and behaviors towards technology. For example, the COVID-19 pandemic increased acceptance of care robots

Examples of methods for simulating future scenarios, ordered by increasing fidelity.

APPROACH	Text vignette	Multi-modal description	Virtual & 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	Here are some potential risks of nanotechnology. Which item is most important to avoid? 1. Losing your personal privacy to tiny new surveillance devices 2. The uncontrollable spread of self-replicating nano-sized robots. 3. …	The Moral Machine: What should the self-driving car do?	Virtual Reality simulation of autonomous vehicle moral dilemmas	Obedience to a robotic boss	Mars Desert Research Station

Increasing fidelity

Figure 1: *Examples of methods for simulating future scenarios, ordered by increasing fidelity.*

[39], and shifts in religiosity can affect attitudes towards assisted reproduction technologies [40]. Unexpected changes in social context between the sci-fi-sci experiment and the actual deployment of the technology may thus threaten prospective validity.

These are formidable challenges. Without a set of accepted principles and methods to address them, behavioral 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 behavioral 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 behavioral 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 (see Figure 1). For instance, many studies 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, such as in situations like neuro-enhancement drugs, where the technology cannot be simulated otherwise [41, 42, 43, 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 they are familiar with [45]. In order to limit such unwanted variations, experimenters need to provide as much contextual information 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, 47, 48]. Although these mockups cannot execute the full range of capabilities of the technology they are simulating, they embed an anticipated future technology within a context that participants are accustomed to. For example, Longoni and Cian [47] presented participants with a mobile app that purported to provide recommendations of an AI master chocolatier. Simulated apps may also allow to use immersive behavioral 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 autonomous vehicles were initially studied through vignettes with static images [23], then later with interactive computer graphics [25], and most recently with virtual reality [49, 50, 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, where human confederates impersonated AI chatbots

with responsive abilities that were at the time impossible for the chatbots to generate themselves [52]. 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 cameras [53]. Finally, at the far end of the fidelity spectrum, some research made use of extremely elaborate staged settings, inaccessible to most researchers—for example, the Mars missions simulations [54, 55, 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, while the core of sci-fi-sci revolves around experimental studies involving human participants, we recognize the valuable role that computational methods such as Monte Carlo simulations and agent-based modeling can play in enriching our understanding of the future. These methods allow researchers to model complex systems and explore how assumptions about human behavior and technological adoption might play out over time. In the same spirit, AI simulations using bots programmed with different behavioral 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? But the solution space also has considerable room to grow. As noted, these methods are scattered across various disciplines, from psychol-

ogy to experiential ethnographic futures [57, 58] that rarely find themselves in conversation. Encouraging discussion among the diverse researchers keen on testing behavioral 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 comport to 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 the science fiction novel *Too Like the Lightning,* author Ada Palmer [59] describes a future where the hyper-mobility afforded by free, flying, supersonic self-driving cars led to the collapse of nation states, replaced by global communities of like-minded individuals. While this technology makes for a fantastic story, it would arguably be inadequate as the focus of a sci-fi-sci experiment. To start with, the development of such technology is very far from our current capabilities. This makes it difficult to imagine its plausible specifications, and even harder to envision the societal landscape by the time it might be invented. In addition, 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 behavior 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 sim-

ply too large to conduct a valid experiment. This concept is often captured visually via a so-called *futures cone* [60], which we have adapted for our purposes (Figure 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 behavior (e.g. cognitive archaeology [61], historical psychology [60] and economic history [62]) as well as future behavior (e.g. 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 up for debate, of course. But one place to start is by studying the near-term behavioral effects of applications which 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. Behavioral experiments can explore how such capabilities may disrupt existing social dynamics—e.g., making people less inhibited in accusing others of lying [63]. When it comes to technologies that do not yet exist, one heuristic is to consider the Technological Readiness Level (TRL) of the target technology [64]. 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 not well formulated, to level 9, where the technology is validated in its intended operational environment. While 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 will 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 time scale of the effects 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 time scale beyond that of a behavioral experiment. For example, we could study the social
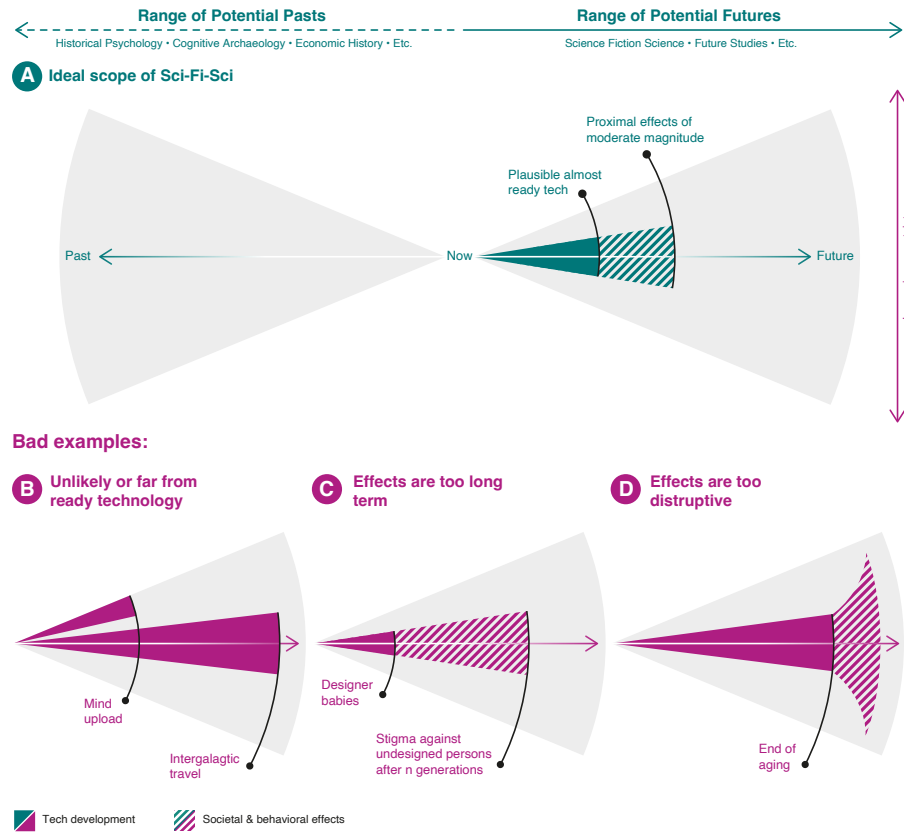
Futures cone for Science Fiction Science

Figure 2: *A modified 'Futures' cone [60] for Science Fiction Science. (**A**) The ideal domain for sci-fi-sci is to study proximal, moderate magnitude effects of plausible, almost ready technology. (**B–D**) show different deviations from this ideal: Topics less amenable to sci-fi-sci include unlikely or far from ready technologies (**B**); societal and behavioral effects that only materialize too far into the future after the introduction of the technology (**C**); and technologies whose effects that too disruptive that they would change so many aspects of society all at once (**D**).*

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 (e.g., anti-aging 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. While science fiction writers may naturally gravitate toward 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' (e.g., the Internet) and 'revolutionary' (e.g., steam power, 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 like AI, which will almost certainly have 'very strong' and possibly 'revolutionary' impact. Sci-fi-sci can and should engage with AI, yet 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 tech-

nology 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 behavioral impacts can be credibly tested in a laboratory context.

While 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 backcasting, 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 behavioral 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 who are already studying the behavior of future humans, in disparate fields, 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 behavior of future humans interacting with future technology in a future social world raises unusual challenges for behavioral scientists, which call for unconventional methods. Institutional mechanisms that can help overcome these challenges include: higher risk tolerance among research funding organizations to create programs 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, so as 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 science fiction science will depend on the quality of its methods.

Contribution Statement

The conceptualisation and writing of this article was a joint effort of all three authors.

Competing Interests

The authors declare no competing interests.

Acknowledgments

References

[1] Braghieri, L., Levy, R. & Makarin, A. Social media and mental health. *American Economic Review* **112**, 3660–3693 (2022).

[2] Van Bavel, J. J., Robertson, C. E., Del Rosario, K., Rasmussen, J. & Rathje, S. Social media and morality. *Annual Review of Psychology* **75** (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 and Chemical toxicology* **42**, 1181–1193 (2004).

[5] Siegrist, M. & Hartmann, C. Consumer acceptance of novel food technologies. *Nature 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. *Perspectives on Psychological Science* **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. *Nature Human Behaviour* **3**, 251–256 (2019).

[8] Cobb, M. D. & Macoubrie, J. Public perceptions about nanotechnology: Risks, benefits and trust. *Journal of Nanoparticle Research* **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. *Science communication* **27**, 240–267 (2005).

[10] Danaher, J. & Sætra, H. S. Mechanisms of techno-moral change: A taxonomy and overview. *Ethical theory and moral practice* **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. *Journal of political Economy* **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. *The quarterly journal of economics* **121**, 289–320 (2006).

[13] Baker, R. *Before bioethics* (Oxford University Press, 2013).

[14] Shariff, A., Green, J. & Jettinghoff, W. The privacy mismatch: Evolved intuitions in a digital world. *Current Directions in Psychological Science* **30**, 159–166 (2021).

[15] Zuboff, S. *The age of surveillance capitalism: The fight for a human future at the new frontier of power* (NY: PublicAffairs, 2019).

[16] Collingridge, D. *The Social Control of Technology* (Pinter, London, 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 and Information Technology* **24**, 35 (2022).

[20] Hopster, J. K. & Maas, M. M. The technology triad: disruptive AI, regulatory gaps and value change. *AI and Ethics* 1–19 (2023).

[21] Brey, P. *Ethics of emerging technology*, 175–191 (Rowman & Littlefield International, 2017).

[22] Pohl, F. Editorial: The great new inventions. *Galaxy Magazine* **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. *Nature Human Behaviour* **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, by 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 (2020).

[27] Luetge, C. The german ethics code for automated and connected driving. *Philosophy & Technology* **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 and Information Technology* **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. *Transportation Research Interdisciplinary Perspectives* **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. *Annals of Tourism Research* **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. *Technology in Society* 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. *Nature Machine Intelligence* **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. *Social Media+ Society* **5**, 2056305119859294 (2019).

[37] Munger, K. Temporal validity as meta-science. *Research & 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. *Advances in experimental social psychology* **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. *Technology in Society* **76**, 102464 (2024).

[40] Inhorn, M. C. & Birenbaum-Carmeli, D. Assisted reproductive technologies and culture change. *Annual Review of Anthropology* **37**, 177–196 (2008).

[41] Dinh, C. T., Humphries, S. & Chatterjee, A. Public opinion on cognitive enhancement varies across different situations. *American Journal of 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. *Consciousness and Cognition* **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 Communications* **4**, 1–14 (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. *Journal of Marketing* **86**, 91–108 (2022).

[48] Köbis, N. *et al.* Artificial intelligence can facilitate selfish decisions by altering the appearance of interaction partners. *arXiv preprint arXiv:2306.04484* (2023).

[49] Benvegnù, 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 (VR)*, 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. *Science and engineering ethics* **25**, 399–418 (2019).

[52] Riek, L. D. Wizard of Oz studies in HRI: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction* **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 (RO-MAN)*, 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, volume 1. *Journal of Futures Studies* **23** (2019).
This and the second volume of the same 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.

[58] Candy, S. & Potter, C. Design and futures, volume 2. *Journal of Futures Studies* **23** (2019).

[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. Cognitive archaeology and the cognitive sciences. *Human paleoneurology* 177–208 (2015).

[62] North, D. C. Structure and performance: The task of economic history. *Journal of Economic literature* **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** (2024).

[64] Héder, M. From NASA to EU: the evolution of the TRL scale in public sector innovation. *The Innovation Journal* **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. *Technological Forecasting and Social Change* **72**, 117–144 (2005).

[66] Goodall, N. J. Ethical decision making during automated vehicle crashes. *Transportation Research Record* **2424**, 58–65 (2014).

[67] Hess, D. J. Incumbent-led transitions and civil society: Autonomous vehicle policy and consumer organizations in the united states. *Technological Forecasting and Social Change* **151**, 119825 (2020).

[68] Kriebitz, A., Max, R. & Lütge, C. The german act on autonomous driving: why ethics still matters. *Philosophy & technology* **35**, 29 (2022).

[69] Mac Síthigh, D. & Siems, M. The Chinese social credit system: A model for other countries? *The Modern Law Review* **82**, 1034–1071 (2019).

[70] Orgad, L. & Reijers, W. A dystopian future? The rise of social credit systems. Tech. Rep. RSCAS 2019/94, Robert Schuman Centre for Advanced Studies (2019).

[71] Tirole, J. Digital dystopia. *American Economic Review* **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] European Commission. 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` (2021). Accessed 10 January 2024.

[74] Turley, P. *et al.* Problems with using polygenic scores to select embryos. *New England Journal of Medicine* **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. *Journal of Law and the Biosciences* **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. Why ethics matters for autonomous cars. *Autonomous driving: Technical, legal and social aspects* 69–85 (2016).

[83] Milakis, D. Long-term implications of automated vehicles: An introduction. *Transport Reviews* **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. *Transport reviews* **39**, 29–49 (2019).

[85] Commission, E., Centre, J. R., Fernández Llorca, D. & Gómez, E. *Trustworthy autonomous vehicles – Assessment criteria for trustworthy AI in the autonomous driving domain* (Publications Office of the European Union, 2021).

[86] Asimov, I. *I, robot* (Gnome Press, 1950).

[87] McDermott, D. What matters to a machine? *Machine ethics* 88–114 (2011).

[88] Hirai, K., Hirose, M., Haikawa, Y. & Takenaka, T. The development of honda humanoid robot. In *Proceedings. 1998 IEEE international conference on robotics and automation (Cat. No. 98CH36146)*, vol. 2, 1321–1326 (IEEE, 1998).

[89] Ishiguro, H. *et al.* Robovie: an interactive humanoid robot. *Industrial robot: An international journal* **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. *Applied Artificial Intelligence* **14**, 145–189 (2000).

[92] Reijers, H. A. Business process management: The evolution of a discipline. *Computers in Industry* **126**, 103404 (2021).

[93] Fortune Business Insights. 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. Tech. Rep. (2024).

[94] Dafoe, A. *et al.* Cooperative ai: machines must learn to find common ground. *Nature* (2021).

[95] Acemoglu, D. & Restrepo, P. Automation and new tasks: How technology displaces and reinstates labor. *Journal of economic perspectives* **33**, 3–30 (2019).

[96] Crandall, J. W. *et al.* Cooperating with machines. *Nature communications* **9**, 233 (2018).

[97] Ishowo-Oloko, F. *et al.* Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence* **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** (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. *Scientific Reports* **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. *Nature Communications* **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. *arXiv preprint arXiv:2303.12712* (2023).

[103] Wu, Q. *et al.* Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155* (2023).

[104] Bamberger, S., Clark, N., Ramachandran, S. & Sokolova, V. How generative ai is already transforming customer service. *Boston Consulting Group* (2023).