nature human behaviour

Perspective

https://doi.org/10.1038/s41562-023-01742-2

Machine culture

Received: 22 August 2023

Accepted: 3 October 2023

Published online: 20 November 2023

Check for updates

Levin Brinkmann[1,11], Fabian Baumann[1,11], Jean-François Bonnefon[2,11], Maxime Derex[2,3,11], Thomas F. Müller[1,11], Anne-Marie Nussberger[1,11], Agnieszka Czaplicka[1], Alberto Acerbi[4], Thomas L. Griffiths[5], Joseph Henrich[6], Joel Z. Leibo[7], Richard McElreath[8], Pierre-Yves Oudeyer[9], Jonathan Stray[10] & Iyad Rahwan[1,11]

The ability of humans to create and disseminate culture is often credited as the single most important factor of our success as a species. In this Perspective, we explore the notion of 'machine culture', culture mediated or generated by machines. We argue that intelligent machines simultaneously transform the cultural evolutionary processes of variation, transmission and selection. Recommender algorithms are altering social learning dynamics. Chatbots are forming a new mode of cultural transmission, serving as cultural models. Furthermore, intelligent machines are evolving as contributors in generating cultural traits—from game strategies and visual art to scientific results. We provide a conceptual framework for studying the present and anticipated future impact of machines on cultural evolution, and present a research agenda for the study of machine culture.

The ability of humans to create and disseminate culture is considered the single most important factor in our species' dominance on Earth[1]. The evolution of human culture has been the subject of extensive study in all of the behavioural sciences, including anthropology[1], psychology[2], cognitive science[3], biology[4], linguistics[5,6], archaeology[7], sociology[8] and economics[9] (Box 1).

Cultural evolution exhibits key Darwinian properties. Culture has been shown to exhibit variation, transmission and selection, and it evolves through the selective retention of cultural traits as well as non-selective processes such as drift[10]. Major shifts in any of these three Darwinian properties can greatly impact cultural evolution. For instance, between 1300 and 1600, European culture experienced successive major shifts due to increased exposure to Chinese technology such as gunpowder, which changed the nature of warfare (variation)[11]; Gutenberg's invention of the printing press (transmission)[12]; and renewed interest in Classical ideas and values, such as Classical ideals of artistic expression, during the Renaissance (selection)[13]. When such substantial changes co-occur, they induce rapid and major impacts on culture.

While new technologies have always affected the course of cultural evolution, in this Perspective we argue that intelligent machines will exert a transformative influence on cultural evolution through their impact on all three Darwinian properties of culture: variation, transmission and selection (Fig. 1). This process began in the early days of the internet with machine-based content ranking by search engines and social media feed algorithms influencing what information people get from others. More recently, generative algorithms have begun participating in the creation of cultural traits themselves. We are observing not only a transformation of human culture but also its evolution into machine culture—culture mediated or generated by machines. This Perspective aims to provide researchers across disciplines with a primer and a roadmap for navigating this monumental shift. As the impact of an increasingly digital society on cultural evolution has been explored elsewhere[14], we specifically focus on the current and potential impact of intelligent machines on cultural evolution. For the purposes of this Perspective, we use the terms 'intelligent machines' and 'artificial intelligence (AI) systems' interchangeably, with AI referring to the

[1]Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany. [2]Toulouse School of Economics, Toulouse, France. [3]Institute for Advanced Study in Toulouse, Toulouse, France. [4]Department of Sociology and Social Research, University of Trento, Trento, Italy. [5]Department of Psychology and Department of Computer Science, Princeton University, Princeton, NJ, USA. [6]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA. [7]DeepMind Technologies Ltd, London, UK. [8]Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [9]Inria, Flowers team, Université de Bordeaux, Bordeaux, France. [10]Center for Human-Compatible Artificial Intelligence, University of California, Berkeley, Berkeley, CA, USA. [11]These authors contributed equally: Levin Brinkmann, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, Thomas F. Müller, Anne-Marie Nussberger, Iyad Rahwan. e-mail: brinkmann@mpib-berlin.mpg.de; rahwan@mpib-berlin.mpg.de

BOX 1

Glossary

Culture: information capable of affecting individuals' behaviours that is acquired from other individuals via social transmission.

Cultural evolution: the change of cultural information over time; the key properties for an evolutionary process are variation, transmission and selection.

Social learning: learning that is influenced by the observation of another individual or their products.

AI: the science and technology enabling machines to perform tasks that typically require human intelligence, such as perceiving the environment, planning and executing actions, and adapting by learning from data or experience.

Machines: intelligent machines. Used interchangeably with AI systems, thus referring to machines that may possess capabilities to perceive the environment, plan and execute actions, and adapt by learning from data or experience.

Variation: the existence of different cultural traits within a population. It represents the raw material on which other processes, such as selection and transmission, operate. Humans and machines add to existing cultural variation through random and guided exploration, as well as recombination of existing cultural traits.

Transmission: the process by which cultural information, such as knowledge, behaviours, traditions or practices, is passed from one individual to another through social learning mechanisms such as observation or teaching.

Selection: the process by which certain cultural traits, practices or ideas become more or less prevalent in a population over time due to differential adoption.

science and technology that allow machines to perform tasks that typically require human intelligence such as perceiving the environment, planning and executing actions, and adapting by learning from data or experience[15,16].

Examples of machine-mediated cultural evolution

We begin by presenting empirical evidence of machine cultural evolution, setting the stage for a detailed exploration through a framework that discusses instances where machines mediate or generate cultural traits from a cultural evolutionary perspective. Through four pivotal examples, we illustrate the diverse ways intelligent machines are transforming cultural evolutionary dynamics. Generative machines, such as text-to-image algorithms, are contributing to the variety of cultural traits. Models drawing on reinforcement learning are pushing humans onto novel ground—for instance, in the ancient game of Go. Large language models (LLMs) are facilitating the transmission of cultural knowledge and redefining the value of human intellectual skills. Meanwhile, transmission pathways are rewired by recommender systems selecting what and from whom humans learn. At first glance, the examples provided might seem to pertain to vastly different technological areas and to translate into a collection of unrelated effects. However, even now, machines are beginning to integrate features from a range of the outlined technologies—reinforcement learning, for instance, is enhancing generative AI. Furthermore, these technologies are operating on multiple levels; generative AI not only generates novel ideas but also offers recommendations for their refinement.

Cultural recombination through generative AI

Generative AI has seen two major waves of innovation in recent years. The inception of generative adversarial networks by Goodfellow et al. in 2014 enabled the algorithmic generation of high-fidelity images[17]. Generative adversarial networks offered capabilities beyond the generation of lifelike images—they also have the ability to blend or interpolate, giving birth to novel creations such as fantasy lifeforms[18]. Subsequent advancement in 2022 saw the advent of diffusion-based text-to-image generative AI systems such as DALL·E, Midjourney and Stable Diffusion. These models substantially enhanced the recombination power of the early models by generating high-resolution images conditioned on text descriptions[19–21]. The original DALL·E, although now surpassed by other models in terms of image quality, demonstrated such recombination capabilities impressively[19]. For instance, when prompted to produce "an armchair in the shape of an avocado", it creatively recombined these two distinct concepts (Fig. 2).

These models can thus increase cultural variation by helping humans to produce new and relevant recombinations, which are sometimes recognized as works of art and sold at prestigious auction houses[22]. While recombination often forms the foundation of human creativity[23], it is still debated how, or even whether, machines can generate relevant content beyond the boundaries of human culture. Even simple latent representations can disentangle the semantic meaning of linguistic concepts[24]. Similarly, text-to-image models use language as a cognitive tool to disentangle and subsequently recombine visual concepts[25]. However, these models build on concepts harvested from human culture. As such, text-to-image models may be limited to the concepts defined and demonstrated by humans in the underlying dataset. However, generative AI models can produce novel "future art" by forecasting future art movements[26] and by deliberately avoiding classification into established artistic movements—such creations have been evaluated as more creative than prestigious contemporary works by human artists[27]. When applied to engineering, similar methods can lead to the discovery of designs that are both novel and superior[28].

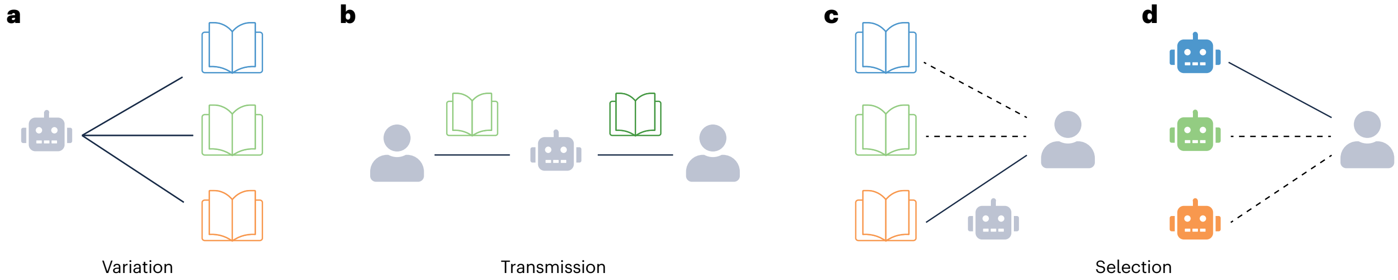

Fig. 1 | Examples of machine culture. a, Novel cultural artefacts are generated through machines. **b**, Machines transmit and potentially mutate cultural artefacts. **c**, Machines select between different cultural artefacts. **d**, Humans select among diverse machines.

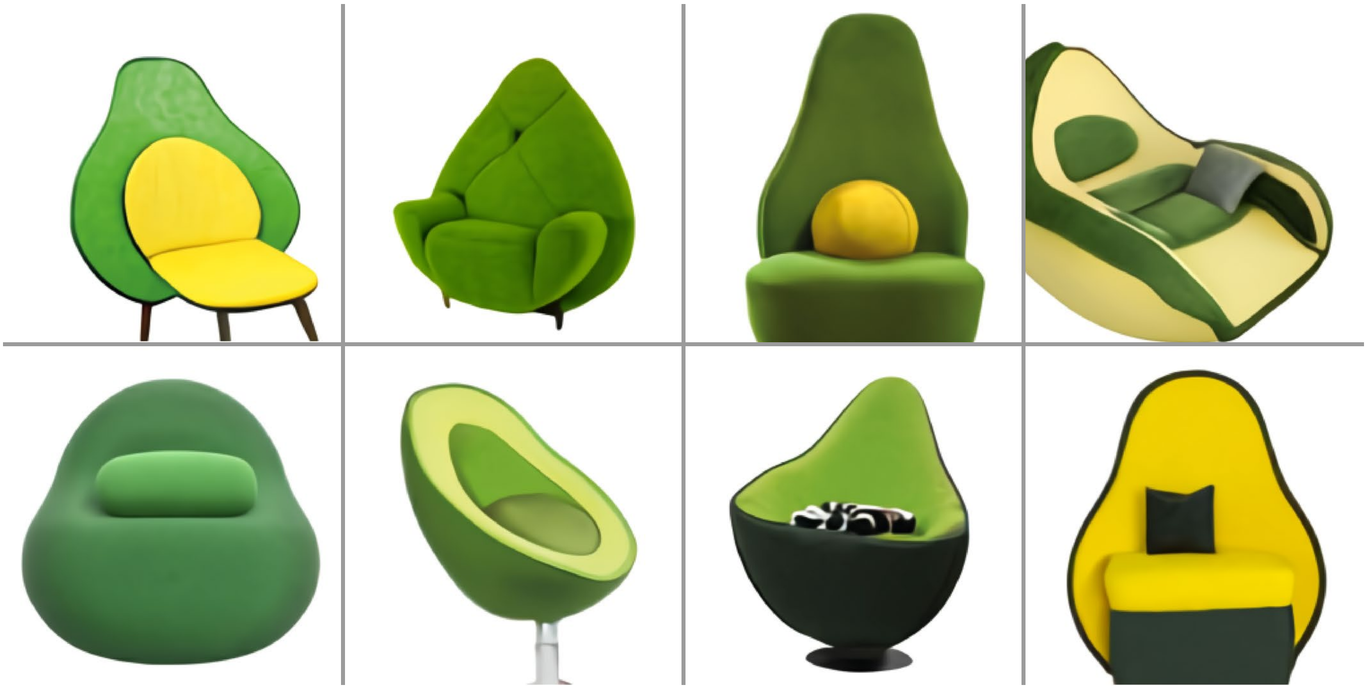

Fig. 2 | Recombination of visual concepts. The avocado chair, synthesized by OpenAI's text-to-image generative AI, DALL·E, exemplifies the early stages of algorithmic cultural recombination. By seamlessly combining previously learned concepts—avocados and chairs in this case—the model showcases the ability to create coherent integrations of disparate elements and demonstrates novelty through recombination. Reproduced from ref. 194.

Generative AI exemplifies the dual nature of machines as both cultural artefacts and creators thereof. On the one hand, generative algorithms are increasingly presented and, to some extent, recognized as authors of art[22]. Simultaneously, machines are subjected to cultural processes of comparison, distribution, modification and eventual abandonment.

Cultural innovation through reinforcement learning

In 2016, AlphaGo defeated Lee Sedol, the world champion Go player, with a series of four victories over five games. Remarkably, AlphaGo managed to surprise Sedol with distinctively non-human gameplay. In particular, move 37 in the second game was considered extremely unconventional, estimated by AlphaGo itself to have a 1 in 10,000 chance of being made by a human[29] (Fig. 3a). AlphaGo's unconventional gameplay probably originated in its self-play training: while its training started with reproducing human gameplay from a large database, AlphaGo also trained through self-play, selecting promising but uncertain moves and evaluating their success against itself[30]. It thereby iteratively improved and developed novel game strategies. It was one of these innovative, self-trained strategies that took Sedol by surprise. As it turned out, it was not even necessary for the model to start from learning human gameplay at all. The successor of AlphaGo, named AlphaGo Zero, ignored all of the accumulated Go knowledge of humankind[30]. Starting from a blank slate, it not only rediscovered human Go strategies but also developed strategies that surpassed those of its human creators.

The innovations generated by AlphaGo and AlphaGo Zero soon entered human culture, as shown by research comparing human gameplay before and after the algorithms' introduction[31]. The decision quality, as measured by an open-source variant of AlphaGo Zero, showed very little improvement in human gameplay from 1950 to 2016, followed by a sudden improvement after the introduction of AlphaGo in March 2016 (refs. 32,33) (Fig. 3b). However, this improvement was not solely due to humans adopting strategies developed by AlphaGo. It also reflected an unexpected shift, wherein humans started developing moves that were qualitatively distinct both from previous human moves and from the novel moves introduced by AlphaGo. In summary, AlphaGo served as an early, quantifiable exemplar of machine culture, generating novel cultural variations through genuine, non-human innovation. This was followed by a major transition into an even broader range of traits as the result of humans building on the previous discoveries made by machines. As the methods underpinning AlphaGo have been generalized to other games and extended to scientific problems, we anticipate a continued infusion of machine-generated discoveries across diverse domains of human culture[34,35].

Language models transmit and revalue cultural knowledge

The release of ChatGPT, a widely accessible LLM, has revolutionized how we interact with machines, using them to learn, brainstorm and refine ideas. Trained on extensive human text data, both historical and contemporary, LLMs act as models of human culture[25], facilitating cultural transmission across individuals and generations. In due course, students began requesting LLMs to complete their homework[36], knowledge workers used LLMs (at their own peril) to automatically extract and summarize required content[37], and software developers widely adopted LLMs as powerful code-writing assistants[38]. LLMs not only serve as content creators but also, for better or worse, act as reservoirs of knowledge and providers of learning opportunities. ChatGPT thereby exemplifies social learning from machines.

As the capabilities and uses of LLMs continue to develop, the value of certain human skills will shift. Some skills may lose value quickly, especially in language-related and cognitively demanding occupations such as translation, copywriting or proofreading[39]. Occupations with more creative uses of language may follow, given that LLMs may soon surpass the creativity of humans as measured by standardized tests[40]. Even though not all occupations will be affected in the short term, a recent survey projects that around 20% of the workforce will experience LLMs impacting at least half of their tasks[39]. It is important to note, however, that this projection is an early estimate and as such inherently uncertain in its accuracy and reliability. Meanwhile, other skills may

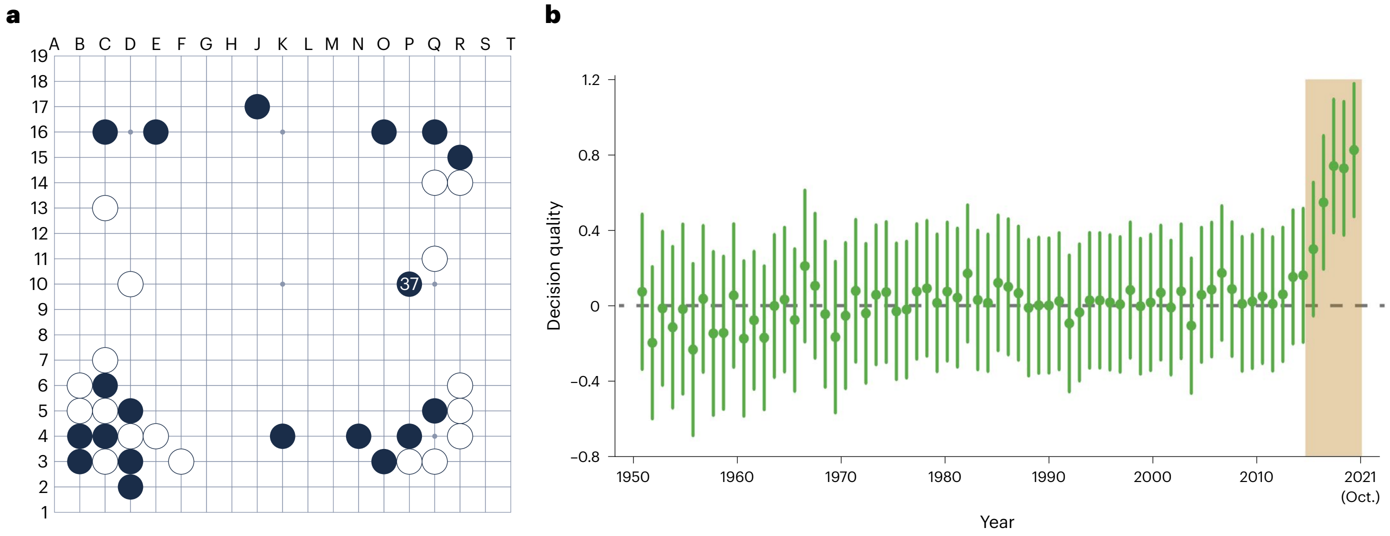

Fig. 3 | Go play before and after the introduction of AlphaGo. a, AlphaGo, in its match against Go world champion Lee Sedol, made a highly unusual and strategic 37th move by placing its stone further from the edge, towards the centre of the board, deviating from the traditional strategy of securing territory along the periphery during the early stages of the game. With this unconventional move, AlphaGo not only broke with centuries-old Go traditions but also paved the way for its ultimate victory in the match. **b**, Decision quality of professional Go players as evaluated by an algorithm performing at a superhuman level. Decision quality significantly increased after Sedol was beaten by AlphaGo on 15 March 2016 (shaded area). Created using data and code from ref. 31.

gain in value—for instance, skills that allow efficient collaboration with LLMs, such as prompt engineering (that is, the skilful writing of instructions to get LLMs to do what we want)[41]. Consequently, human creativity is experiencing a remarkable shift[42]. It is not merely focused on generating final outputs but is increasingly evolving towards interactions with machines, progressing from explicit prompts to more natural conversations[43]. Workers who invest in related skills may outperform workers who do not, potentially accelerating the adoption of LLMs and further increasing the value of knowing how to collaborate with them.

Cultural rewiring through recommender systems

The digital age is so data-rich that it has become increasingly hard for humans to navigate available information. In this abundance, recommender systems that manage and filter information have a silent but increasingly important role that is easily overlooked given how seamlessly these systems have integrated into our everyday digital lives. These systems select and prioritize items on the basis of a variety of explicit and implicit features, including personal interests, control settings, past user behaviour and the behaviour of similar users. While recommender systems do not add variation in cultural traits, they demonstrably impact the selective retention and transmission of cultural traits.

By promoting new social ties—such as suggesting whom to follow on X (formerly Twitter), date on Tinder or work with on LinkedIn—recommender systems alter the set of people to whom we are exposed, ultimately changing the structure of our social networks and hence pathways of cultural transmission[44] (Fig. 4a). But recommender systems can also bypass network structures, exerting an even more direct impact on which cultural content or products we are exposed to. For example, e-commerce websites and streaming platforms deploy recommender systems to steer customers through the expansive array of available products on the basis of content and collaborative filtering. Content filtering matches information about a customer's consumption history with the attributes of all available products to make suggestions about related purchases: a customer who has recently purchased running shoes may receive suggestions for additional running equipment, matched to the shoes in price and design[45]. Meanwhile, collaborative filtering[46] makes recommendations based on less obvious patterns of correlations in users' profiles: if users A and B overlap in their previously consumed movies and music, the recommender system might suggest content of a completely different kind—for example, recommending to user A a book that user B liked (Fig. 4b). In sum, recommender systems influence cultural evolution by rewiring our social networks and modifying information flows such that they can substantially influence the dynamics of cultural markets[47–49].

A framework for machine-mediated cultural evolution

Building on these exemplary instances of how machine technologies may impact cultural evolution, we will now map out a systematic framework for studying the potential of machines to shape cultural evolutionary processes.

Culture has been defined as information capable of affecting individuals' behaviours that is acquired from other individuals via social transmission[50]. Consequently, the science of cultural evolution examines the change of cultural information over time[51,52]. Cultural information is represented by individual cultural traits, which can exist as cognitive representations or be expressed in behaviours or artefacts[53]. Culture evolves according to a process similar to the one by which species change—that is, through the selective retention of cultural traits and through other non-selective processes, such as drift[10]. While there are ongoing discussions on how far the analogy between cultural and genetic evolution should be pushed[54,55], there is general agreement that culture exhibits the key properties of evolutionary systems: variation, transmission and selection. These properties are not necessarily the result of distinct processes of human cognition and behaviour, yet they offer a useful framework to analyse the multifaceted ways machines can influence cultural evolution (see Table 1 for a summary).

Variation refers to the existence of different cultural traits within a population. Transmission involves the spread of cultural information from one individual to another through social learning mechanisms, including observation and teaching. During this transmission, information losses often occur, affecting the preservation of cultural traits. Selection occurs when certain cultural traits are more likely to be adopted by individuals due to factors such as their usefulness,

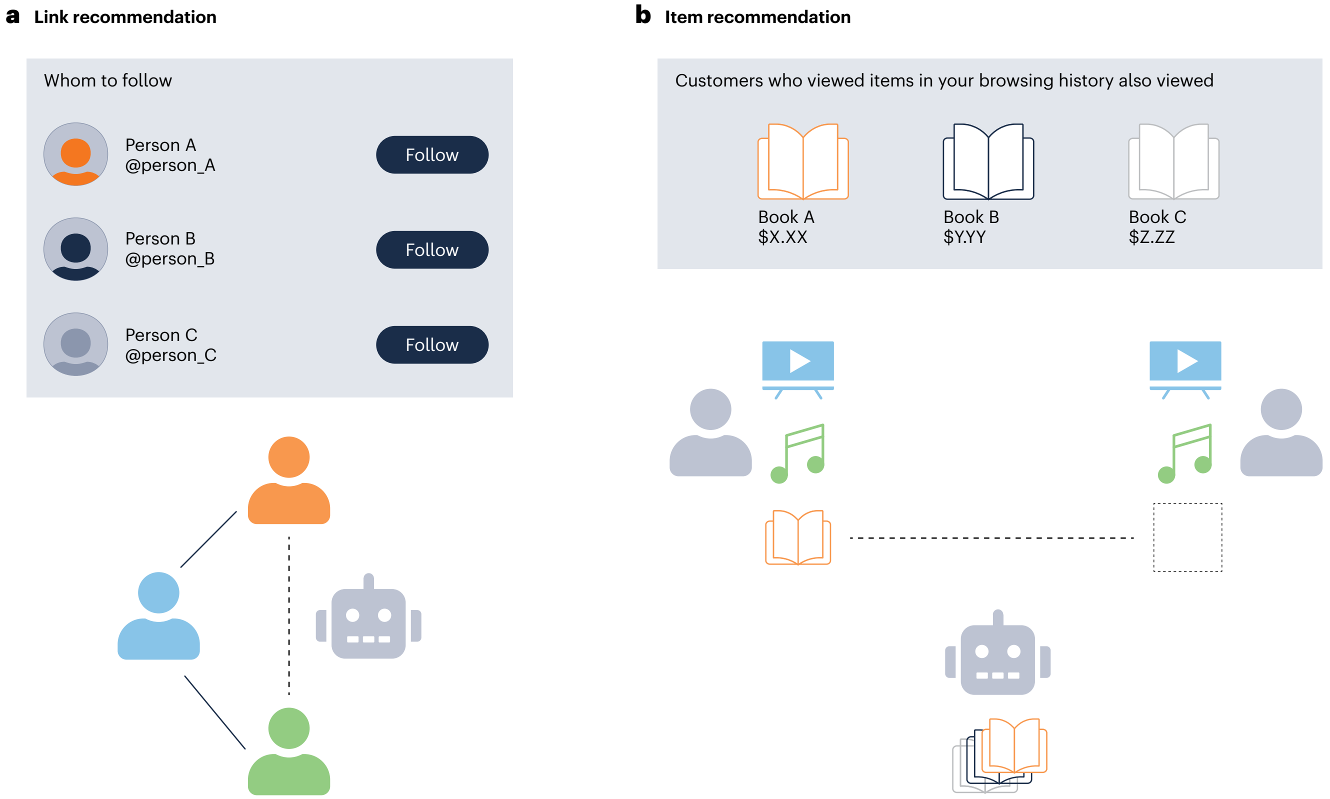

Fig. 4 | Exemplary instances of cultural rewiring. a, Friendship recommendation on an online platform (for example, X or LinkedIn). A new social tie (person A) is recommended to the green user. **b**, Collaborative filtering for item recommendations. The recommender system suggests a book to the right user on the basis of the correlations in co-purchases of other items (movie and music album) by the left user. The top panels represent the user interfaces; the bottom panels provide a schematic depiction of the underlying mechanism.

popularity or compatibility with existing cultural practices. Crucially, the prevalence of specific traits over time can be influenced by both their selection and variations in transmission fidelity, contingent on the traits' characteristics. We anticipate that machine technologies will affect each of these three properties (Fig. 1), and these changes are likely to have transformative impacts on cultural evolution.

Variation

Variation refers to the presence of diverse cultural traits within a population. Humans and machines contribute to this existing cultural variation through both random and guided exploration, as well as the recombination of existing cultural traits. However, machines, leveraging their unique capacities, can produce traits distinct from those produced by humans, thus potentially steering culture towards new paths.

Machines have the ability to learn individually at an unparalleled scale, enabling the discovery of novel cultural traits through extensive exploration. Thorough exploration is not unique to machines. Thomas Edison, for example, famously tested over 6,000 materials to find the most suitable filament for his incandescent light bulb[56]. However, the computational speed of humans imposes time constraints that limit their exploration compared with machines[57]. For instance, AlphaGo Zero needed only three days to play 4.9 million games against itself, achieving a superhuman level of proficiency[30]. No human can amass that volume of experience individually; hence, we tend to rely heavily on pre-existing, culturally evolved solutions that are socially learned[58,59]. This approach narrows the scope of innovation to the cultural context of previous generations. AI systems such as AlphaGo use reinforcement learning—based on iterative rounds of trial and error—and hence can transcend this cultural path dependency by starting from a blank slate and discovering innovative solutions through exploration alone[60]. As a result, AI systems have the potential to generate culturally alien traits. In this respect, AlphaGo is in stark contrast to language models, which primarily learn to reproduce human cultural output.

Machines and humans employ guiding models—policies—that aid in efficiently exploring complex solution spaces, thereby enabling the discovery of solutions that would be unobtainable through random exploration alone. For instance, Mesoamerican skywatchers, via an early example of human astronomic modelling, developed a policy for optimal seed planting times, maximizing agricultural productivity[61]. Similar to humans, algorithms such as AlphaGo use complex policies to guide their search. The general approach to guided search is similar for both humans and machines: experience is used to update a predictive model of the world, which then informs actions that generate new experiences. Machines no longer require humans to explicitly formulate these guiding models. Instead, general-purpose neural networks are universal approximators, allowing functional relationships to be learned[62,63]. This allows for unprecedented model complexity and consequently pushes the boundary of attainable solutions[64]. Surpassing the capabilities of any human-conceived model, the neural-network-based algorithm AlphaFold has achieved remarkable precision in modelling intramolecular relations, enabling it to tackle one of molecular biology's greatest challenges: predicting a protein's three-dimensional structure solely on the basis of its DNA sequence[65]. For decades, protein structure prediction has drawn substantial intellectual contributions from scientists, even harnessing the collective intelligence of tens of thousands of citizen scientists[66]. Nevertheless, AlphaFold's complex model of intramolecular relations, manifested in the form of a deep neural network, proved superior, eclipsing these extensive human efforts with remarkable proficiency.

Table 1 | Tentative conjectures about ways in which machines might shape the processes of cultural evolution

Process	Machine capability	Possible impact on culture
Variation	Learning at unprecedented scale and speed (for example, reinforcement learning)	Emergence of solutions culturally alien to humans
	Superhuman model complexity	Generation of solutions inconceivable by collective human intelligence and human cultural evolution
	Incomparably broad and deep knowledge base	Creation of novel recombinations beyond the human horizon
	(Semi-)autonomous creativity (for example, image generation models)	Facilitation or crowding out of human participation in creative exploration
Transmission	Exposing and/or preserving documented cultural knowledge (for example, LLM chatbots)	Enhanced retrieval and increased transmission fidelity of documented cultural knowledge
	Reproduction of human biases	Amplification or mitigation of existing biases; potential for increased cultural erosion
	Accelerated processing of empirical evidence	Facilitation of transmission of less compressed knowledge
	Leveraging the unique cognitive capabilities of both humans and machines	Expanding the collective capacity to maintain diverse cultural artefacts
Selection	Recommendation of social ties (for example, link recommendation algorithm)	Shaping social networks, potentially inhibiting or enhancing serendipitous encounters
	Curating content (for example, ranking algorithms and collaborative filtering)	Indirectly shaping social networks via content exposure; shaping incentives for human content creators (for example, clickbait)
	Adapting to human feedback (for example, RLHF)	Alignment of machines to human goals, potentially leading to unintended consequences (for example, spread of highly believable myths)
	Machine learning from machine-generated content	Selection of machine-generated content as main driver of cultural evolution
	Adaptability of AI models to market/societal demands	Proliferation of appealing apps with varying alignment to human welfare
	Competing with humans in cultural production (for example, poetry)	Specialization of humans and machines in distinct niches

Many contemporary algorithms, such as LLMs, have been trained on a cultural repertoire of unprecedented scale[67]. For instance, some LLMs are trained on hundreds of billions of words[68]. While size does not guarantee diversity[69], the degree to which machines can learn from human cultural repertoires far exceeds what can be achieved by a single human being, allowing, for instance, individual language models to cover a broad spectrum of languages comprehensively[70]. By combining knowledge from disjointed communities—spanning cultural and geographic contexts, language barriers or scientific disciplines—intelligent machines can produce novel recombinations that may be beyond humanity's reach. Irrespective of whether LLMs truly exhibit understanding[71], the mere recombination of cultural traits can lead to innovation and fuel cultural evolution[69,72]. Indeed, recombination is considered central to generating novel cultural traits[23,69,73–75]. Another way in which intelligent machines may increase the variety of cultural traits is by building on human knowledge while at the same time identifying and deliberately avoiding common human pathways. For instance, in the realm of scientific discoveries, intelligent machines may be designed to uncover scientifically plausible and promising 'alien' hypotheses that would normally fall outside the focal point of contemporary scientific communities[76]. This does not imply, however, that intelligent machines are without their own limitations or blind spots, as we will discuss later.

But machines may also augment human exploration. Conditional generative AI, such as DALL·E, allows users to steer the generative process through text prompts, enabling experimentation with diverse concepts and visualizations without the need for advanced artistic skills[19]. Throughout human history, the effective size of the population that contributed to the creation and maintenance of cultural knowledge has constrained human cultural evolution[1]. Technology has had a varying impact on human cultural activities. On the one hand, it can democratize behaviour. For instance, advancements in digital cameras and editing software have enabled individuals with limited technical knowledge to capture high-quality photographs and edit them effectively. On the other hand, the advent of new technologies can lead to increasingly specialized roles for humans. For instance, visual effects technology has led to the emergence of entirely new professions in film production, contributing to a cumulative increase in crew size[77]. Lastly, technological advancements can also lead to a reduction in human engagement in certain activities. For example, the advent of farming decreased the necessity for hunting, as people could grow food instead. It is yet to be seen whether the proliferation of increasingly intelligent machines will follow a similar trajectory. Will it democratize and increase human participation in creative fields, will it lead to new professions and increasing specialization, or could it lead to a decline in human-led creative expression, as we become increasingly reliant on intelligent machines[42]? Undoubtedly, human cultural exploration, in science as in arts, will increasingly be in collaboration with machines[78,79].

Transmission

The transmission of cultural information occurs via social learning, defined as learning that is influenced by the observation of other individuals and/or their products[58]. This is in contrast to individual learning, where information is acquired, for instance, through trial and error. In humans, social learning tends to be imperfect, which means that transmission events are associated with a risk of losing cultural information. For example, if individuals observe a behaviour incorrectly or do not fully understand it, they may not pass on the correct information to others. Over time, this can lead to the deterioration or even loss of cultural practices, which can have lasting consequences for a population.

Intelligent machines will increasingly be involved in the preservation and transmission of cultural information. Cultural evolution has supplied humans with increasingly efficient tools to preserve cultural information. The invention of writing, for instance, allowed humans to mitigate cultural loss by recording information in a more permanent way. Theoretical and empirical studies of cultural evolution have shown that the stability of cultural information strongly depends on both the size of the population that shares information[73] and the type of social learning mechanisms involved (with lower fidelity when transmission relies on mere observation than when it relies on verbal teaching, for instance)[80]. Besides serving as a persistent medium of cultural storage analogous to a book, machines can learn to seek and transmit information[79,81] and can act as conversational and pedagogical agents, similar to teachers[82]. This dual role has potential for drastically boosting cultural preservation by reducing cultural drift. For instance, LLMs have been harnessed to resurrect historical figures[83] and to revive languages teetering on the edge of extinction[84]. By fostering interactive cultural experiences for learners, these technologies can enhance the comprehension and retention of cultural information[85].

As machines store and transmit cultural information, they may reproduce and transmit biases inherent to human culture (for example, through biased training datasets); but they also offer potential to mitigate those biases. Machine learning models reproduce various content biases inherent to their training data, including gender bias, racial or ethnic bias, bias for negative and threat-related information, and socio-economic bias[69,86–90]. For instance, many LLMs make implicit assumptions about the gender associated with professions such as nurse and doctor, reflecting the demographic perspective dominant in the training data[91]. When an LLM is used for the revision and enhancement of human-written text, it can transmit and/or reinforce these biases[69]. Numerous strategies exist to address bias and fairness in machine learning models[92]. For instance, the demographic representation in the outputs of LLMs can be improved by using prompts featuring personas from a broad demographic spectrum[93]. However, LLMs present another challenge: they struggle to accurately represent languages and communities for which there are limited training data[69,94]. This limitation implies that LLM-based copy-editing could inadvertently lead to cultural erosion within these underrepresented communities due to losses in text reproduction. Nevertheless, LLMs, with their capacity to handle vast quantities of training data, can contribute to the preservation and even enhancement of cultural diversity, provided that the training data are carefully curated with attention to diversity and representation.

Cultural transmission can be impacted not only by the replication of biases but also by disparities between human and algorithmic biases. Examples of biases found in humans are confirmation bias[95], availability bias[96] and a bias towards specific symmetries[97]. Despite their reputation for undermining optimal decision-making, biases can actually reflect optimal decisions within a particular socio-environmental context under cognitive constraints such as memory, computation or experience[98–102]. Similarly, machines may seize biases—for instance, deep learning architectures assume spatial symmetries to improve training efficiency[95]. Importantly, humans and machines operate under different cognitive constraints[57] and inhabit distinct socio-environmental contexts. This may give rise to idiosyncratic biases. For instance, due to their enhanced computational capabilities, machines may display more utilitarian rationality[91]. Remarkably, humans often anticipate more utilitarian behaviour from them[103] and might indeed, as we will elaborate in the following section, enhance this tendency. Irrespective of their origin, the effects of biases on culture tend to strengthen through repeated transmission[99,104,105]. This can restrict the spectrum of solutions that a population can derive, and even highly effective solutions may not be sustained if they conflict with pre-existing biases[106]. As such, the increased cognitive diversity—for instance, with regard to biases—within human–machine societies has the potential to expand the collective capacity to maintain diverse cultural artefacts. Artefacts that might not be maintained by humans could be maintained by machines. Conversely, in the transmission between humans and machines, a misalignment of biases can increase the risk of information loss[107].

Machines' increased computational capacity might additionally affect the feasibility of accumulating uncompressed information via a 'big data' approach. The compressibility of information is the inverse of its algorithmic complexity: compressing information is achieved by creating a rule that is shorter than a complete list of the data itself[108]. Compression is a key feature of both human cognition and machine learning[109]; any information-processing system must address a general trade-off between a truthful representation of the raw information and constraints on computation. The extent and nature of these constraints, however, differ between humans and machines[57]. Compression processes have an important role in human transmission to mitigate cultural loss: raw information often is not learnable because it is too complex, whereas a compressed rule might very well be learnable. Human language, for example, retains its expressivity by becoming learnable via the evolution of compressed structure[110]. Similarly, scientists develop, transmit and revise theories as compressed representations of knowledge. Machine learning has the potential to reduce some of the constraints resulting from human computation, as the amount of data that can be processed is vastly increased. Consequently, information might be increasingly transmitted with low compression—in the form of big data—when predictive power is of ultimate importance, and for some applications, the necessity to derive and transmit highly compressed rules or theories might be reduced[111,112]. For instance, with the advent of AlphaFold, scientists might focus on collecting and preserving further ground-truth data on molecule structures to refine future models rather than refining and transmitting theories of atomic interactions. While symbolic representations may remain crucial for efficient computation, algorithmic-assisted discovery could lessen the need for their transmission, as these representations could be easily regenerated[113]. However, theories may retain importance in shaping human understanding and intuition, serving as essential tools for conceptualizing knowledge—a function that the framework we present in this work aims to fulfil.

Selection

Culture evolves in part through the selective retention of cultural traits. In the context of machine culture, selection can occur at a level where machines select what and from whom humans learn, at a level where humans select machine behaviour, and at a level where there is selection between humans and machines.

Regarding what humans learn, social learning strategies shape what, when and whom we copy. These strategies can be broadly categorized into content-based and context-based strategies[114,115]. Content-based strategies consider what is learned, favouring, for instance, social over non-social information[116]. Conversely, context-based strategies attend to situational features, focusing on properties of a cultural model (for example, their competence, success, prestige, knowledge or similarity), frequencies (for example, most common behaviour or rare behaviour) or internal states of the learner (for example, uncertainty or the cost of individual learning). Machines that help humans to navigate vast information spaces by (pre-)selecting cultural traits often reflect such social learning strategies.

For instance, content-based filtering algorithms aim to maximize the similarity between items a user previously showed interest in and unobserved items. Emulating context-based strategies in selection, ranking algorithms typically sort items according to a relevance score, which is based on the items' popularity[117,118]. Concurrently, collaborative filtering algorithms detect hidden patterns between items and users, achieving recommendations about novel items to similar users without using additional exogenous information about individual items or users[119] (Fig. 3b).

Another dimension along which algorithms may influence the selective retention of cultural traits pertains to social networks, which form the backbone of information exchange. In this context, social ties are rewired as users follow algorithmic recommendations based on user attributes, such as popularity, or similarities in user preferences in both personal (X: "Who to follow") and professional domains (LinkedIn: "People you might know"). Link recommendation algorithms have the potential to shape the overall evolution of social networks[120–122]. X's "who to follow" recommendation was observed to disproportionately benefit those users who were already the most popular, fuelling "the rich get richer" dynamics[122]. A growing body of research documents a complex but persistent and critical relationship between social networks' structure and collectives' ability to collaborate, coordinate and solve problems[123–125] that ultimately shape cultural repertoires[74,126].

While machines have most commonly relied on exploiting explicit user preferences and historical behaviour (for example, ratings and engagement), there has been a growing interest in considering users' internal states to improve algorithmic recommendations. For instance, users might be uncertain about their preferences—especially in domains in which they lack expertise. Bayesian approaches can model users' uncertainty and be used to update algorithmic recommendations as the user interacts with the system[127]. As another example, recommender systems might account for the cognitive cost of exploring items or learning about them by prioritizing items that are easier for users to evaluate or learn[128]. Affective recommender systems[129] use techniques such as natural language processing to make inferences about users' emotional states and even combine them with other context information such as the recommendation domain (for example, music or movies)[130–132]. While recommender systems have so far been mostly shaping user preferences implicitly (for example, by optimizing the position or ranking of content), LLMs may accelerate developments where users are increasingly persuaded explicitly through interactive argumentation.

Downstream consequences of selection by machines may often be specific to particular environments, algorithmic models and feedback loops. However, one feature generalizing across various contexts is that algorithms—by the design of underlying business models—are often geared towards maximizing user engagement for profit[133]. In social networks, this may be achieved by promoting content congruent with users' past engagement or ingroup attitudes[134], or content that humans inherently attend to, such as emotionally and morally charged content[135,136]. One example is information that relates to threat or elicits disgust, as shown in transmission chain experiments inspired from cultural evolutionary theory[137]. The algorithmic amplification of such content may then feed back into human social learning—for instance, inflating beliefs about the normative value of expressing moral outrage[138,139], increasing outgroup animosity[140] or creating echo chambers and filter bubbles[141–143]. It is important to note that user engagement is a signal of value to both users and platforms deploying algorithms, connecting them in complex feedback loops[144,145]: machines such as recommender systems react to user engagement, selecting types of content that people engage with. Users also react to recommender systems, both directly in terms of clicking, viewing and purchasing and in terms of what they produce, as content creators anticipate what will receive the widest distribution. These feedback loops, but also deliberate product design choices along with policy approaches, provide potent leverage points for aligning recommender systems with human values[146]. A promising approach to addressing this challenge could involve considering potential misalignments between users' engagement and their own preferences and identifying the boundary conditions determining when maximizing user engagement enhances user welfare and when it produces the opposite effect[147]. Algorithmic systems more generally offer powerful ways to bridge social divides—for instance, by designing selection policies that steer users' attention to content that increases mutual understanding and trust[148,149], or by identifying and promoting links in social networks that can effectively mitigate polarizing dynamics[150,151]. Machine selection can also be deliberately geared towards fostering content diversity[152] or towards maximizing agreement among humans with diverse preferences[153].

Human preferences, in turn, can directly shape machine behaviour, in particular through reinforcement learning with human feedback (RLHF)[154,155]. One way to conceptualize this is by viewing machines as students that generate arrays of solutions, with humans acting as teachers who select the most suitable ones. This process can nudge machines towards desirable properties such as helpfulness, honesty and harmlessness[155]. However, harvesting human annotators' preferences at scale through RLHF can induce machines to adopt behaviours unintended by the deploying organizations, exemplified by chatbots that increasingly endorse inflated political views or express a heightened desire to avoid shutdown[156]. Human influence on machine behaviour also occurs through more subtle pathways: through training on human text alone, LLMs picked up a tendency to repeat users' stated views[156]. Machines' attention to human feedback may create selection pressures towards pleasing human interlocutors. For instance, humans may favour machines catering to non-factual stories and narratives that match concepts and ideas preferred by human cognition[157], such as those appealing to intuitive expectations about the natural world[158,159] or to specific religious practices[160].

Humans select machine behaviour also through direct creation and curation of training data for machine learning, and through more indirect interactions with machine-generated outputs. Despite efforts to watermark machine-created content[161], machine-made and human-made content will increasingly intertwine[162]. For instance, it is estimated that many supposedly human-written texts on crowdsourcing platforms are already augmented by machines[163]. It therefore seems inevitable that future machines will be trained on mixed human–machine content, forming part of a larger feedback cycle between content generation and selection. The human element in this cycle may prove crucial—for instance, in preventing 'model collapse', a dynamic where repeated training on machine-generated data narrows its outputs to very few traits[162]. This phenomenon stems from a classification bias that favours more prevalent classes[164], a bias that is amplified through successive iterations of learning[165]. By contrast, when encountering similar challenges, human culture might preserve diversity by utilizing biases with counteracting effects, such as endorsing local conformity[166]. That said, it seems likely that machines could recover similar strategies to maintain cultural diversity without human involvement.

Yet another pathway for human selection on machine behaviour pertains to general machine properties. For instance, humans may choose between different intelligent machines available on the market on the basis of factors such as preferences, cost, usefulness, harmfulness and alignment with regulatory requirements. As such, the language model LLaMA recently gained attention for its relatively smaller size, making it more cost-effective to use[167]. Conversely, ChatGPT outperformed many comparable models due to its superior accessibility and helpfulness[155]. Human demands towards machines are likely to change over time, shaping the trajectory of machine culture similar to other historic cases where technologies evolved in response to changing human demands (for example, the wheel from wooden wheels on carriages to rubber tires on automobiles). Crucially, if intelligent machines are designed and evaluated by non-representative experts, these systems run the risk of unintentionally reproducing and intensifying the biases inherent to their selectors[168].

Selection is also bound to occur at a level where machines or humans are favoured over one another. For instance, the ability of machines to process vast amounts of information both quickly and accurately affords them a competitive edge in numerous cognitive tasks, such as strategic gameplay and information retrieval. Relatedly, due to their cost-effectiveness and efficiency, intelligent machines may grow into the main workforce across various professional domains[39,169]. However, analogous to how the invention of the car did not diminish interest in running as a sport, the proliferation of machines may not curtail human interest in intellectual pursuits; instead, it might redirect the focus from necessity to leisure and entertainment. Meanwhile, across various contexts, humans might favour other humans over machines due to their shared experience. For instance, even though present-day machines can conceive messages perceived as more empathic than messages conceived by humans[170,171], such messages may be perceived differently once recipients become aware of their artificial origin. This phenomenon, which could be referred to as the "artificial empathy paradox"[172], may, at least in part, arise from the very fact that empathizing is effortful for humans[173] and that, as such, it conveys a motivational social signal to others that is made void by machine involvement. Selection between humans and machines does not imply that one agent dominates over the other in any cultural niche: often, one will augment rather than fully replace the other; at other times, the presence of the other may trigger the development of new skills or roles.

Grand challenges and open questions

We now suggest a broad research agenda for computational and behavioural scientists interested in the phenomenon of machine culture.

Measurement

One major open challenge is to quantify how much of human cultural dynamics can be attributable to algorithmic processes. For instance, it is difficult to completely disentangle the effects of ranking and recommendation algorithms on culture from alternative processes of human social learning, such as communication technology, institutions and social practices. Since the inception of human culture, derived tools have had an important role in shaping cultural processes, making the establishment of a baseline a challenging question in itself. Getting good estimates is a precondition to optimizing for the usefulness of these algorithms while avoiding undesirable cultural impacts such as polarization[174]. This challenge is reflected in research on "filter bubbles", which has moved from considering algorithmic curation as decisive force shaping online engagement[142] towards acknowledging the influence of users' own choices on social media ranking algorithms[175] and search engines[176], thus highlighting the intricate feedback loops between machine and human decisions. While we hope that it will be possible to find appropriate metrics for the cultural impact of machines, the intricacy of this problem qualifies it as an open grand challenge.

A complementary question is how to quantify the influence of machine-generated artefacts—for example, artwork, literature and music—on human cultural production in these areas. As generative AI becomes more commonplace, distinguishing intrinsic human culture from machine-generated culture or machine-influenced culture becomes even more challenging, especially as watermarking techniques may not be universally adopted. Despite popular media claims about machine-generated art developing its own unique style[177], we do not yet have a reliable way of verifying these claims, let alone assessing machine-to-human cultural transmission[107].

Another measurement concern relates to the quantification of the cultural regularities encoded in LLMs and other AI models. Even prior to the rise of LLMs, social media platforms such as Facebook already possessed fairly detailed quantitative models of cultural regularities and differences[178], developed mainly for the purpose of marketing to particular demographics. However, as LLMs are trained from curated datasets and fine-tuned using human feedback, quantifying the biases that are introduced and/or mitigated by these models is crucial[179,180].

Societal decision-making

We currently observe a rapid increase in the diversity of AI models, including LLMs, accelerated by the open-source software movement. However, market forces, such as regulation and market power, may result in a world dominated by a small number of monolithic models. This raises the possibility of reduction in cultural diversity, as major social, political and economic forces try to shape global machine culture to match their preferences. This process may be amplified by feedback loops, in which LLMs train on an ongoing basis from synthetic data or from human data that contains much machine-generated text. Preliminary evidence points to the possibility of model collapse, with the models losing diversity and converging to a state with low variance[162].

Conversely, we face a potential 'Tower of Babel' scenario. As AI models become increasingly personalized, conforming to and reinforcing our individual worldviews, they risk engendering an unprecedented fragmentation of our shared perception of the world. In the biblical story (Genesis 11:1–9), the construction of the tower led to a divine intervention that scattered humanity and confounded languages. Drawing a parallel, if we continually interact with machines that echo and affirm our preconceived notions, we risk isolating ourselves within ideologically and culturally homogenous echo chambers. Such fragmentation can stifle meaningful dialogue, breed misunderstanding and, ultimately, fracture our shared future vision.

Against this background, a key research agenda is to quantify the degree to which a given AI model or ecosystem of models exhibits uniformity or diversity. It is also imperative to understand what constitutes a 'healthy' level of diversity, one that retains local sovereignty while also fostering collective human flourishing.

Ensuring that AI models, such as LLMs, reflect the beliefs and values of a given community requires mechanisms for societal decision-making about what knowledge goes into the models[181]. Furthermore, humans exhibit variation in their ethical expectations towards machines, both within and across cultures[182]. This raises questions about how to best aggregate diverse, potentially conflicting preferences to arrive at an agreeable outcome[183]. A number of interesting ideas are emerging, from voting on different algorithmic policymakers[184] to using LLMs to summarize diverse human opinions[185] and generate consensus statements[153].

A related challenge is how to ensure long-term monitoring of machine culture. Similar to the notion of human-in-the-loop control of intelligent machines, we can aspire towards society-in-the-loop control of the complex phenomena of machine culture[186].

Suppose a community knows which cultural beliefs and values it wants to encode in an AI model. The next question is how to ensure that these are indeed present in the model. One approach is to carefully curate the dataset on which the model is trained[187,188]. Another, increasingly used approach is fine-tuning based on RLHF[189]. Yet another approach is to fine-tune using constitutional rules provided by humans[190]. The relative merits and drawbacks of these various approaches are still not well understood. There is a need for methods to check which cultural beliefs and values have been learned, inferred or encoded in a given AI model and the degree to which they align with a target culture.

Long-term dynamics and optimization

In all likelihood, the future of culture will be hybrid, with cultural artefacts—scientific theories, industrial processes, art and literature—being created by a combination of human and machine intelligence. This raises a suite of open questions relating to the long-term dynamics of human–machine co-evolution. These dynamics may lead to diverse phenomena ranging from different forms of human–machine mutualism to Red Queen effects (an evolutionary arms race between humans and machines) that characterize the co-evolution of both forms of intelligence[191].

A related question is how to optimize the aforementioned dynamics, in order to combine human and machine intelligence in an ideal or safe manner[76]. This may be relevant to questions of risk mitigation. Although experts disagree on the timescales and the degree of risk involved, the potential of superhuman artificial general intelligence poses a possible existential threat to the human species[192]. Cultural evolution provides a useful framework for navigating this challenge.

Specifically, cultural evolution processes take place today at multiple scales, with human collectives—for example, companies, universities, institutions, cities and nation states—acting as the units of selection[193]. This multi-level selection can, in principle, operate at the level of human organizations augmented by intelligent machines and (eventually) superhuman artificial general intelligence. Engineering this evolutionary process can provide means for ensuring human survival and agency in the long run.

Conclusion

We asked GPT-4 to first write a compressed version of this Perspective and then to provide a conclusion. It suggested the following (minimal editing to align nomenclature has been applied). The symbiosis of human and machine intelligence is forging a new epoch of cultural evolution. This Perspective highlights the transformative role of intelligent machines in reshaping creativity, redefining skill value and altering human interactions. Central to the discourse is the triad of cultural evolution: variation, transmission and selection, and how machines interface with each. The interaction is multifaceted, from generative AI birthing novel cultural artefacts to recommendation algorithms influencing individual perspectives. However, the crux remains in understanding and navigating the challenges and opportunities that arise from this hybridization of culture. As the imprints of intelligent machines grow deeper, it is imperative to ensure a harmonious co-creation of culture where both humans and machines augment, rather than eclipse, each other. This will not only broaden the horizons of cultural exploration but also fortify the tapestry of human experience in the age of intelligent machines.

References

1. Henrich, J. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter* (Princeton Univ. Press, 2016).
2. Heyes, C. *Cognitive Gadgets: The Cultural Evolution of Thinking* (Harvard Univ. Press, 2018).
3. Thompson, B., van Opheusden, B., Sumers, T. & Griffiths, T. L. Complex cognitive algorithms preserved by selective social learning in experimental populations. *Science* **376**, 95–98 (2022).
4. Whiten, A. Cultural evolution in animals. *Annu. Rev. Ecol. Evol. Syst.* **50**, 27–48 (2019).
5. Gray, R. D. & Atkinson, Q. D. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature* **426**, 435–439 (2003).
6. Kirby, S., Cornish, H. & Smith, K. Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language. *Proc. Natl Acad. Sci. USA* **105**, 10681–10686 (2008).
7. Shennan, S. *Genes, Memes, and Human History: Darwinian Archaeology and Cultural Evolution* (Thames & Hudson, 2002).
8. Kiley, K. & Vaisey, S. Measuring stability and change in personal culture using panel data. *Am. Sociol. Rev.* **85**, 477–506 (2020).
9. Mokyr, J. *A Culture of Growth: The Origins of the Modern Economy* (Princeton Univ. Press, 2017).
10. Mesoudi, A., Whiten, A. & Laland, K. N. Perspective: is human cultural evolution Darwinian? Evidence reviewed from the perspective of the origin of species. *Evolution* **58**, 1–11 (2004).
11. Needham, J. in *Chemistry and Chemical Technology, Pt. 7: Military Technology—the Gunpowder Epic* Vol. 5 (Cambridge Univ. Press, 1986).
12. Eisenstein, E. L. *The Printing Press as an Agent of Change* Vol. 1 (Cambridge Univ. Press, 1980).
13. Mesoudi, A. Culture and the Darwinian Renaissance in the social sciences and humanities: for a special issue of the *Journal of Evolutionary Psychology*, "The Darwinian Renaissance in the Social Sciences and Humanities". *J. Evol. Psychol.* **9**, 109–124 (2011).
14. Acerbi, A. *Cultural Evolution in the Digital Age* (Oxford Univ. Press, 2019).
15. Russell, S. & Norvig, P. *Artificial Intelligence: A Modern Approach* (Prentice Hall, 2009).
16. Kurzweil, R., Richter, R., Kurzweil, R. & Schneider, M. L. *The Age of Intelligent Machines* (MIT Press, 1990).
17. Goodfellow, I. et al. Generative adversarial networks. *Commun. ACM* **63**, 139–144 (2020).
18. Epstein, Z., Boulais, O., Gordon, S. & Groh, M. Interpolating GANs to scaffold autotelic creativity. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2007.11119 (2020).
19. Ramesh, A. et al. Zero-shot text-to-image generation. In *International Conf. on Machine Learning* 8821–8831 (PMLR, 2021).
20. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with CLIP latents. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2204.06125 (2022).
21. Rombach, R. et al. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.* 10684–10695 (2022).
22. Epstein, Z., Levine, S., Rand, D. G. & Rahwan, I. Who gets credit for AI-generated art? *iScience* **23**, 101515 (2020).
23. Thagard, P. & Stewart, T. C. The AHA! experience: creativity through emergent binding in neural networks. *Cogn. Sci.* **35**, 1–33 (2011).
24. Mikolov, T., Yih, W. & Zweig, G. Linguistic regularities in continuous space word representations. In *Proc. 2013 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 746–751 (Association for Computational Linguistics, 2013).
25. Colas, C., Karch, T., Moulin-Frier, C. & Oudeyer, P.-Y. Language and culture internalization for human-like autotelic AI. *Nat. Mach. Intell.* **4**, 1068–1076 (2022).
26. Lisi, E., Malekzadeh, M., Haddadi, H., Lau, F.D.-H. & Flaxman, S. Modelling and forecasting art movements with CGANs. *R. Soc. Open Sci.* **7**, 191569 (2020).
27. Elgammal, A., Liu, B., Elhoseiny, M. & Mazzone, M. CAN: Creative Adversarial Networks, generating 'art' by learning about styles and deviating from style norms. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1706.07068 (2017).
28. Wang, Y., Shimada, K. & Barati Farimani, A. Airfoil GAN: encoding and synthesizing airfoils for aerodynamic shape optimization. *J. Comput. Des. Eng.* **10**, 1350–1362 (2023).
29. Metz, C. In two moves, AlphaGo and Lee Sedol redefined the future. *Wired* (16 March 2016).
30. Silver, D. et al. Mastering the game of Go without human knowledge. *Nature* **550**, 354–359 (2017).
31. Shin, M., Kim, J., van Opheusden, B. & Griffiths, T. L. Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Proc. Natl Acad. Sci. USA* **120**, e2214840120 (2023).
32. Choi, S., Kim, N., Kim, J. & Kang, H. How does AI improve human decision-making? Evidence from the AI-powered Go program. Preprint at *SSRN* https://doi.org/10.2139/ssrn.3893835 (2022).
33. Shin, M., Kim, J. & Kim, M. Human learning from artificial intelligence: evidence from human Go players' decisions after AlphaGo. *Proc. Annu. Meet. Cogn. Sci. Soc.* **43**, 43 (2021).
34. Schrittwieser, J. et al. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature* **588**, 604–609 (2020).
35. Fawzi, A. et al. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature* **610**, 47–53 (2022).
36. Kasneci, E. et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn. Individ. Differ.* **103**, 102274 (2023).
37. Wagner, G., Lukyanenko, R. & Paré, G. Artificial intelligence and the conduct of literature reviews. *J. Inf. Technol.* **37**, 209–226 (2022).

38. Chen, M. et al. Evaluating large language models trained on code. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2107.03374 (2021).
39. Eloundou, T., Manning, S., Mishkin, P. & Rock, D. GPTs are GPTs: an early look at the labor market impact potential of large language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2303.10130 (2023).
40. Stevenson, C., Smal, I., Baas, M., Grasman, R. & van der Maas, H. Putting GPT-3's creativity to the (alternative uses) test. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2206.08932 (2022).
41. Popli, N. How to get a six-figure job as an AI prompt engineer. *Time* https://time.com/6272103/ai-prompt-engineer-job/ (14 April 2023).
42. Epstein, Z., Hertzmann, A. & the Investigators of Human Creativity. Art and the science of generative AI. *Science* **380**, 1110–1111 (2023).
43. Oppenlaender, J. The creativity of text-to-image generation. In *Proc. 25th International Academic Mindtrek Conference* 192–202 (Association for Computing Machinery, 2022); https://doi.org/10.1145/3569219.3569352
44. Li, Z. (L.), Fang, X. & Sheng, O. R. L. A survey of link recommendation for social networks: methods, theoretical foundations, and future research directions. *ACM Trans. Manage. Inf. Syst.* **9**, 1–26 (2018).
45. Lops, P., de Gemmis, M. & Semeraro, G. in *Recommender Systems Handbook* (eds Ricci, F. et al.) 73–105 (Springer US, 2011); https://doi.org/10.1007/978-0-387-85820-3_3
46. Su, X. & Khoshgoftaar, T. M. A survey of collaborative filtering techniques. *Adv. Artif. Intell.* **2009**, 421425 (2009).
47. Anderson, A., Maystre, L., Anderson, I., Mehrotra, R. & Lalmas, M. Algorithmic effects on the diversity of consumption on Spotify. In *Proc. Web Conference 2020* 2155–2165 (Association for Computing Machinery, 2020).
48. Krumme, C., Cebrian, M., Pickard, G. & Pentland, S. Quantifying social influence in an online cultural market. *PLoS ONE* **7**, e33785 (2012).
49. 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).
50. Richerson, P. J. & Boyd, R. *Not by Genes Alone: How Culture Transformed Human Evolution* (Univ. of Chicago Press, 2005).
51. Cavalli-Sforza, L. L. & Feldman, M. W. *Cultural Transmission and Evolution: A Quantitative Approach* (Princeton Univ. Press, 1981).
52. Mesoudi, A. Pursuing Darwin's curious parallel: prospects for a science of cultural evolution. *Proc. Natl Acad. Sci. USA* **114**, 7853–7860 (2017).
53. Enquist, M. & Ghirlanda, S. Evolution of social learning does not explain the origin of human cumulative culture. *J. Theor. Biol.* **246**, 129–135 (2007).
54. Acerbi, A. & Mesoudi, A. If we are all cultural Darwinians what's the fuss about? Clarifying recent disagreements in the field of cultural evolution. *Biol. Phil.* **30**, 481–503 (2015).
55. Morin, O. Reasons to be fussy about cultural evolution. *Biol. Phil.* **31**, 447–458 (2016).
56. Weitzman, M. L. Recombinant growth. *Q. J. Econ.* **113**, 331–360 (1998).
57. Griffiths, T. L. Understanding human intelligence through human limitations. *Trends Cogn. Sci.* **24**, 873–883 (2020).
58. Boyd, R. & Richerson, P. J. *Culture and the Evolutionary Process* (Univ. of Chicago Press, 1985).
59. Mesoudi, A. *Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences* (Univ. of Chicago Press, 2011).
60. Leibo, J. Z., Hughes, E., Lanctot, M. & Graepel, T. Autocurricula and the emergence of innovation from social interaction: a manifesto for multi-agent intelligence research. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1903.00742 (2019).
61. Aveni, A. F. *Skywatchers: A Revised and Updated Version of Skywatchers of Ancient Mexico* (Univ. of Texas Press, 2001).
62. Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **4**, 251–257 (1991).
63. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
64. Zenil, H. et al. The future of fundamental science led by generative closed-loop artificial intelligence. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2307.07522 (2023).
65. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
66. Cooper, S. et al. Predicting protein structures with a multiplayer online game. *Nature* **466**, 756–760 (2010).
67. Bommasani, R. et al. On the opportunities and risks of foundation models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2108.07258 (2022).
68. Hoffmann, J. et al. Training compute-optimal large language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2203.15556 (2022).
69. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: can language models be too big? In *Proc. 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623 (Association for Computing Machinery, 2021).
70. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
71. Mitchell, M. & Krakauer, D. C. The debate over understanding in AI's large language models. *Proc. Natl Acad. Sci. USA* **120**, e2215907120 (2023).
72. Charbonneau, M. Modularity and recombination in technological evolution. *Phil. Technol.* **29**, 373–392 (2016).
73. Henrich, J. Demography and cultural evolution: how adaptive cultural processes can produce maladaptive losses—the Tasmanian case. *Am. Antiq.* **69**, 197–214 (2004).
74. Henrich, J. & Muthukrishna, M. What makes us smart? *Top. Cogn. Sci.* https://doi.org/10.1111/tops.12656 (2023).
75. Youn, H., Strumsky, D., Bettencourt, L. M. A. & Lobo, J. Invention as a combinatorial process: evidence from US patents. *J. R. Soc. Interface* **12**, 20150272 (2015).
76. Sourati, J. & Evans, J. A. Accelerating science with human-aware artificial intelligence. *Nat. Hum. Behav.* https://doi.org/10.1038/s41562-023-01648-z (2023).
77. Tinits, P. & Sobchuk, O. Open-ended cumulative cultural evolution of Hollywood film crews. *Evol. Hum. Sci.* **2**, e26 (2020).
78. Grizou, J., Points, L. J., Sharma, A. & Cronin, L. A curious formulation robot enables the discovery of a novel protocell behavior. *Sci. Adv.* **6**, eaay4237 (2020).
79. Kramer, S., Cerrato, M., Džeroski, S. & King, R. Automated scientific discovery: from equation discovery to autonomous discovery systems. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2305.02251 (2023).
80. Lucas, A. J. et al. The value of teaching increases with tool complexity in cumulative cultural evolution. *Proc. R. Soc. B* **287**, 20201885 (2020).
81. Borsa, D., Piot, B., Munos, R. & Pietquin, O. Observational learning by reinforcement learning. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1706.06617 (2017).
82. Kohnke, L., Moorhouse, B. L. & Zou, D. ChatGPT for language teaching and learning. *RELC J.* **54**, 537–550 (2023).
83. Haller, E. & Rebedea, T. Designing a chat-bot that simulates an historical figure. In *2013 19th International Conference on Control Systems and Computer Science* 582–589 (IEEE, 2013).
84. Zhang, S., Frey, B. & Bansal, M. How can NLP help revitalize endangered languages? A case study and roadmap for the Cherokee language. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2204.11909 (2022).

85. Ijaz, K., Bogdanovych, A. & Trescak, T. Virtual worlds vs books and videos in history education. *Interact. Learn. Environ.* **25**, 904–929 (2017).
86. Buolamwini, J. & Gebru, T. Gender shades: intersectional accuracy disparities in commercial gender classification. In *Proc. 1st Conference on Fairness, Accountability and Transparency* 77–91 (PMLR, 2018).
87. Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
88. O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown, 2016).
89. Prates, M. O., Avelar, P. H. & Lamb, L. C. Assessing gender bias in machine translation: a case study with google translate. *Neural Comput. Appl.* **32**, 6363–6381 (2020).
90. Acerbi, A. & Stubbersfield, J. Large language models show human-like content biases in transmission chain experiments. Preprint at *OSF* https://doi.org/10.31219/osf.io/8zg4d (2023).
91. Vig, J. et al. Investigating gender bias in language models using causal mediation analysis. *Adv. Neural Inf. Process. Syst.* **33**, 12388–12401 (2020).
92. Pessach, D. & Shmueli, E. A review on fairness in machine learning. *ACM Comput. Surv.* **55**, 1–51 (2022). 44.
93. Argyle, L. P. et al. Out of one, many: using language models to simulate human samples. *Political Anal.* **31**, 337–351 (2023).
94. Hendy, A. et al. How good are GPT models at machine translation? A comprehensive evaluation. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2302.09210 (2023).
95. Bartlett, F. C. *Remembering: A Study in Experimental and Social Psychology* xix, 317 (Cambridge Univ. Press, 1932).
96. Kashima, Y. Maintaining cultural stereotypes in the serial reproduction of narratives. *Pers. Soc. Psychol. Bull.* **26**, 594–604 (2000).
97. Griffiths, T. L., Christian, B. R. & Kalish, M. L. Using category structures to test iterated learning as a method for identifying inductive biases. *Cogn. Sci.* **32**, 68–107 (2008).
98. Lieder, F. & Griffiths, T. L. Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* **43**, e1 (2020).
99. Simon, H. A. in *Utility and Probability* (eds Eatwell, J. et al.) 15–18 (Palgrave Macmillan UK, 1990).
100. Todd, P. M. & Gigerenzer, G. Environments that make us smart: ecological rationality. *Curr. Dir. Psychol. Sci.* **16**, 167–171 (2007).
101. Tversky, A. & Kahneman, D. Judgment under uncertainty: heuristics and biases. *Science* **185**, 1124–1131 (1974).
102. Gershman, S. J., Horvitz, E. J. & Tenenbaum, J. B. Computational rationality: a converging paradigm for intelligence in brains, minds, and machines. *Science* **349**, 273–278 (2015).
103. Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J. & Cusimano, C. Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proc. Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* 117–124 (Association for Computing Machinery, 2015).
104. Griffiths, T. L., Kalish, M. L. & Lewandowsky, S. Theoretical and empirical evidence for the impact of inductive biases on cultural evolution. *Phil. Trans. R. Soc. B* **363**, 3503–3514 (2008).
105. Kirby, S., Dowman, M. & Griffiths, T. L. Innateness and culture in the evolution of language. *Proc. Natl Acad. Sci. USA* **104**, 5241–5245 (2007).
106. Thompson, B. & Griffiths, T. L. Human biases limit cumulative innovation. *Proc. R. Soc. B* **288**, 20202752 (2021).
107. Brinkmann, L. et al. Hybrid social learning in human-algorithm cultural transmission. *Phil. Trans. R. Soc. A* **380**, 20200426 (2022).
108. Tamariz, M. & Kirby, S. Culture: copying, compression, and conventionality. *Cogn. Sci.* **39**, 171–183 (2015).
109. Chater, N. & Vitányi, P. Simplicity: a unifying principle in cognitive science? *Trends Cogn. Sci.* **7**, 19–22 (2003).
110. Kirby, S., Tamariz, M., Cornish, H. & Smith, K. Compression and communication in the cultural evolution of linguistic structure. *Cognition* **141**, 87–102 (2015).
111. Anderson, C. The end of theory: the data deluge makes the scientific method obsolete. *Wired* (23 June 2018).
112. Spinney, L. Are we witnessing the dawn of post-theory science? *Guardian* (9 January 2022).
113. Liu, Z., Madhavan, V. & Tegmark, M. AI Poincaré 2.0: machine learning conservation laws from differential equations. *Phys. Rev. E* **106**, 045307 (2022).
114. Kendal, R. L. et al. Social learning strategies: bridge-building between fields. *Trends Cogn. Sci.* **22**, 651–665 (2018).
115. Henrich, J. & McElreath, R. The evolution of cultural evolution. *Evol. Anthropol.* **12**, 123–135 (2003).
116. Mesoudi, A., Whiten, A. & Dunbar, R. A bias for social information in human cultural transmission. *Br. J. Psychol.* **97**, 405–423 (2006).
117. Sharma, D. K. & Sharma, A. A comparative analysis of web page ranking algorithms. *Int. J. Comput. Sci. Eng.* **2**, 2670–2676 (2010).
118. Duhan, N., Sharma, A. K. & Bhatia, K. K. Page ranking algorithms: a survey. In *2009 IEEE International Advance Computing Conference* 1530–1537 (IEEE, 2009).
119. Koren, Y., Rendle, S. & Bell, R. Advances in collaborative filtering. In *Recommender Systems Handbook* (eds Ricci, F., Rokach, L. & Shapira, B.) 91–142 (Springer US, Boston, MA, 2021).
120. Banihashemi, S. & Abhari, A. Effects of different recommendation algorithms on structure of social networks. In *2021 International Conference on Computational Science and Computational Intelligence (CSCI)* 1395–1400 (IEEE, 2021); https://doi.org/10.1109/CSCI54926.2021.00279
121. Ferrara, A., Espín-Noboa, L., Karimi, F. & Wagner, C. Link recommendations: their impact on network structure and minorities. In *14th ACM Web Science Conference 2022*. 228–238 (Association for Computing Machinery, 2022); https://doi.org/10.1145/3501247.3531583
122. Su, J., Sharma, A. & Goel, S. The effect of recommendations on network structure. In *Proc. 25th International Conference on World Wide Web* 1157–1167 (International World Wide Web Conferences Steering Committee, 2016).
123. Lazer, D. & Friedman, A. The network structure of exploration and exploitation. *Adm. Sci. Q.* **52**, 667–694 (2007).
124. Mason, W. & Watts, D. J. Collaborative learning in networks. *Proc. Natl Acad. Sci. USA* **109**, 764–769 (2012).
125. Woolley, A. W., Aggarwal, I. & Malone, T. W. Collective intelligence and group performance. *Curr. Dir. Psychol. Sci.* **24**, 420–424 (2015).
126. Derex, M. & Boyd, R. Partial connectivity increases cultural accumulation within groups. *Proc. Natl Acad. Sci. USA* **113**, 2982–2987 (2016).
127. Kant, V., Jhalani, T. & Dwivedi, P. Enhanced multi-criteria recommender system based on fuzzy Bayesian approach. *Multimed. Tools Appl.* **77**, 12935–12953 (2018).
128. Bollen, D., Knijnenburg, B. P., Willemsen, M. C. & Graus, M. Understanding choice overload in recommender systems. In *Proc. Fourth ACM Conference on Recommender Systems* 63–70 (Association for Computing Machinery, 2010).
129. Tkalcic, M., Kosir, A. & Tasic, J. Affective recommender systems: the role of emotions in recommender systems. In *The RecSys 2011 Workshops-Decisions@ RecSys 2011 and UCERSTI-2: Human Decision Making in Recommender Systems; User-Centric Evaluation of Recommender Systems and Their Interfaces-2* Vol. 811, 9–13 (CEUR-WS.org, 2011).

130. Gonzalez, G., de la Rosa, J. L., Montaner, M. & Delfin, S. Embedding emotional context in recommender systems. In *2007 IEEE 23rd International Conference on Data Engineering Workshop* 845–852 (IEEE, 2007).
131. Osman, N. A., Mohd Noah, S. A., Darwich, M. & Mohd, M. Integrating contextual sentiment analysis in collaborative recommender systems. *PLoS ONE* **16**, e0248695 (2021).
132. Zheng, Y., Mobasher, B. & Burke, R. D. The role of emotions in context-aware recommendation. *Decis. RecSys* **2013**, 21–28 (2013).
133. Zhang, X., Ferreira, P., Godinho De Matos, M. & Belo, R. Welfare properties of profit maximizing recommender systems: theory and results from a randomized experiment. *MIS Q.* **45**, 1 (2021).
134. Levy, R. Social media, news consumption, and polarization: evidence from a field experiment. *Am. Econ. Rev.* **111**, 831–870 (2021).
135. Brady, W. J., Gantman, A. P. & Van Bavel, J. J. Attentional capture helps explain why moral and emotional content go viral. *J. Exp. Psychol. Gen.* **149**, 746–756 (2020).
136. Brady, W. J., Jackson, J. C., Lindström, B. & Crockett, M. J. Algorithm-mediated social learning in online social networks. *Trends Cogn. Sci.* (in the press).
137. Acerbi, A. Cognitive attraction and online misinformation. *Palgrave Commun.* **5**, 1–7 (2019).
138. Brady, W. J. et al. Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility. *Nat. Hum. Behav.* https://doi.org/10.1038/s41562-023-01582-0 (2023).
139. Brady, W. J. & Crockett, M. J. Norm psychology in the digital age: how social media shapes the cultural evolution of normativity. *Perspect. Psychol. Sci.* https://doi.org/10.1177/17456916231187395 (2023).
140. Milli, S., Carroll, M., Pandey, S., Wang, Y. & Dragan, A. D. Engagement, user satisfaction, and the amplification of divisive content on social media. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2305.16941 (2023).
141. Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W. & Starnini, M. The echo chamber effect on social media. *Proc. Natl Acad. Sci. USA* **118**, e2023301118 (2021).
142. Pariser, E. *The Filter Bubble: What the Internet Is Hiding from You* (Penguin, 2011).
143. Sunstein, C. R. *Republic.com 2.0* (Princeton Univ. Press, 2007).
144. Jiang, R., Chiappa, S., Lattimore, T., György, A. & Kohli, P. Degenerate feedback loops in recommender systems. In *Proc. 2019 AAAI/ACM Conference on AI, Ethics, and Society* 383–390 (ACM, 2019).
145. Pagan, N. et al. A classification of feedback loops and their relation to biases in automated decision-making systems. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2305.06055 (2023).
146. Stray, J. et al. Building human values into recommender systems: an interdisciplinary synthesis. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2207.10192 (2022).
147. Kleinberg, J., Mullainathan, S. & Raghavan, M. The challenge of understanding what users want: inconsistent preferences and engagement optimization. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2202.11776 (2022).
148. Ovadya, A. & Thorburn, L. Bridging systems: open problems for countering destructive divisiveness across ranking, recommenders, and governance. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2301.09976 (2023).
149. Yao, B., Jiang, M., Yang, D. & Hu, J. Empowering LLM-based machine translation with cultural awareness. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2305.14328 (2023).
150. Garimella, K., De Francisci Morales, G., Gionis, A. & Mathioudakis, M. Reducing controversy by connecting opposing views. In *Proc. Tenth ACM International Conference on Web Search and Data Mining* 81–90 (Association for Computing Machinery, 2017).
151. Santos, F. P., Lelkes, Y. & Levin, S. A. Link recommendation algorithms and dynamics of polarization in online social networks. *Proc. Natl Acad. Sci. USA* **118**, e2102141118 (2021).
152. Möller, J., Trilling, D., Helberger, N. & Van Es, B. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Inf. Commun. Soc.* **21**, 959–977 (2018).
153. Bakker, M. et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Adv. Neural Inf. Process. Syst.* **35**, 38176–38189 (2022).
154. Christiano, P. F. et al. Deep reinforcement learning from human preferences. *Adv. Neural Inf. Process. Syst.* **30** (2017).
155. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
156. Perez, E. et al. Discovering language model behaviors with model-written evaluations. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2212.09251 (2022).
157. Claidière, N., Scott-Phillips, T. C. & Sperber, D. How Darwinian is cultural evolution? *Phil. Trans. R. Soc. B* **369**, 20130368 (2014).
158. Blancke, S., Van Breusegem, F., De Jaeger, G., Braeckman, J. & Van Montagu, M. Fatal attraction: the intuitive appeal of GMO opposition. *Trends Plant Sci.* **20**, 414–418 (2015).
159. Miton, H. & Mercier, H. Cognitive obstacles to pro-vaccination beliefs. *Trends Cogn. Sci.* **19**, 633–636 (2015).
160. Poulsen, V. & DeDeo, S. Cognitive attractors and the cultural evolution of religion. In *Proc. of the Annual Meeting of the Cognitive Science Society* **45**, 45 (2023).
161. Kirchenbauer, J. et al. A watermark for large language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2301.10226 (2023).
162. Shumailov, I. et al. The curse of recursion: training on generated data makes models forget. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2305.17493 (2023).
163. Veselovsky, V., Ribeiro, M. H. & West, R. Artificial artificial artificial intelligence: crowd workers widely use large language models for text production tasks. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2306.07899 (2023).
164. Japkowicz, N. & Stephen, S. The class imbalance problem: a systematic study. *Intell. Data Anal.* **6**, 429–449 (2002).
165. Kalish, M. L., Griffiths, T. L. & Lewandowsky, S. Iterated learning: intergenerational knowledge transmission reveals inductive biases. *Psychon. Bull. Rev.* **14**, 288–294 (2007).
166. Axelrod, R. The dissemination of culture: a model with local convergence and global polarization. *J. Confl. Resolut.* **41**, 203–226 (1997).
167. Touvron, H. et al. LLaMA: open and efficient foundation language models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2302.13971 (2023).
168. West, S. M., Whittaker, M. & Crawford, K. *Discriminating Systems: Gender, Race and Power in AI* (AI Now Institute, 2019).
169. Autor, D. H. Why are there still so many jobs? The history and future of workplace automation. *J. Econ. Perspect.* **29**, 3–30 (2015).
170. Ayers, J. W. et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern. Med.* **183**, 589–596 (2023).
171. Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C. & Althoff, T. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nat. Mach. Intell.* **5**, 46–57 (2023).
172. Perry, A. AI will never convey the essence of human empathy. *Nat. Hum. Behav.* https://doi.org/10.1038/s41562-023-01675-w (2023).
173. Weisz, E. & Zaki, J. Motivated empathy: a social neuroscience perspective. *Curr. Opin. Psychol.* **24**, 67–71 (2018).

174. Carroll, M., Hadfield-Menell, D., Russell, S. & Dragan, A. Estimating and penalizing preference shift in recommender systems. In *Proc. 15th ACM Conference on Recommender Systems* 661–667 (Association for Computing Machinery, 2021).
175. Bakshy, E., Messing, S. & Adamic, L. A. Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**, 1130–1132 (2015).
176. Robertson, R. E. et al. Users choose to engage with more partisan news than they are exposed to on Google Search. *Nature* **618**, 342–348 (2023).
177. Art made by artificial intelligence is developing a style of its own. *Economist* (24 May 2023).
178. Obradovich, N. et al. Expanding the measurement of culture with a sample of two billion humans. *J. R. Soc. Interface* **19**, 20220085 (2022).
179. Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl Acad. Sci. USA* **115**, E3635–E3644 (2018).
180. Karjus, A., Solà, M. C., Ohm, T., Ahnert, S. E. & Schich, M. Compression ensembles quantify aesthetic complexity and the evolution of visual art. *EPJ Data Sci.* **12**, 21 (2023).
181. Santy, S., Liang, J. T., Bras, R. L., Reinecke, K. & Sap, M. NLPositionality: characterizing design biases of datasets and models. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2306.01943 (2023).
182. Awad, E. et al. The Moral Machine experiment. *Nature* **563**, 59–64 (2018).
183. Brandt, F., Conitzer, V. & Endriss, U. in *Multiagent Systems* (ed. Weiss, G.) 213–284 (MIT Press, 2012).
184. Koster, R. et al. Human-centred mechanism design with Democratic AI. *Nat. Hum. Behav.* **6**, 1398–1407 (2022).
185. Small, C. T. et al. Opportunities and risks of LLMs for scalable deliberation with Polis. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2306.11932 (2023).
186. Rahwan, I. Society-in-the-loop: programming the algorithmic social contract. *Ethics Inf. Technol.* **20**, 5–14 (2018).
187. Jernite, Y. et al. Data governance in the age of large-scale data-driven language technology. In *2022 ACM Conference on Fairness, Accountability, and Transparency* 2206–2222 (Association for Computing Machinery, 2022); https://doi.org/10.1145/3531146.3534637
188. Laurençon, H. et al. The bigscience roots corpus: a 1.6 tb composite multilingual dataset. *Adv. Neural Inf. Process. Syst.* **35**, 31809–31826 (2022).
189. Ziegler, D. M. et al. Fine-tuning language models from human preferences. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1909.08593 (2020).
190. Bai, Y. et al. Constitutional AI: harmlessness from AI feedback. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2212.08073 (2022).
191. Bergstrom, C. T. & Lachmann, M. The Red King effect: when the slowest runner wins the coevolutionary race. *Proc. Natl Acad. Sci. USA* **100**, 593–598 (2003).
192. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies* (Oxford Univ. Press, 2014).
193. Wilson, D. S. et al. Multilevel cultural evolution: from new theory to practical applications. *Proc. Natl Acad. Sci. USA* **120**, e2218222120 (2023).
194. *DALL·E: Creating Images from Text*, https://openai.com/research/dall-e (OpenAI, 2021).

Acknowledgements

L.B. thanks M. Bakker for useful discussions. We thank B. Supriyatno for supporting the formatting of the manuscript. The conclusion was created by GPT-4 on the basis of a summary (also created by GPT-4) of this manuscript with minimal editing to align nomenclature. J.-F.B. and M.D. acknowledge IAST funding from the French National Research Agency (ANR) under grant no. ANR-17-EURE-0010 (Investissements d'Avenir programme).

Author contributions

Conception: L.B. and I.R. Manuscript preparation: L.B., F.B., J.-F.B., M.D., T.F.M., A.-M.N. and I.R. Critical review, commentary or revision: A.C., A.A., T.L.G., J.H., J.Z.L., R.M., P.-Y.O. and J.S. Supervision: I.R.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Levin Brinkmann or Iyad Rahwan.

Peer review information *Nature Human Behaviour* thanks Olivier Morin, Sune Lehmann and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.