Societal AI: Research Challenges and Opportunities

Microsoft Research Asia

Contents

Foreword by Lidong Zhou 2

Foreword by James A. Evans 4

Preface 6

Introduction to Societal AI 8

Societal AI Research Questions 10

How can AI be aligned with diverse human values and ethical principles? 12

How can AI systems be designed to ensure fairness and inclusiveness across different cultures, regions, and demographic groups? 16

How can we ensure AI systems are safe, reliable, and controllable, especially as they become more autonomous? 18

How can human-AI collaboration be optimized to enhance human abilities? 20

How can we effectively evaluate AI's capability and performance in new, unforeseen tasks and environments? 22

How can we enhance AI interpretability to ensure transparency and trust in its decision-making processes? 24

How will AI reshape human cognition, learning, and creativity, and what new capabilities might it unlock? 26

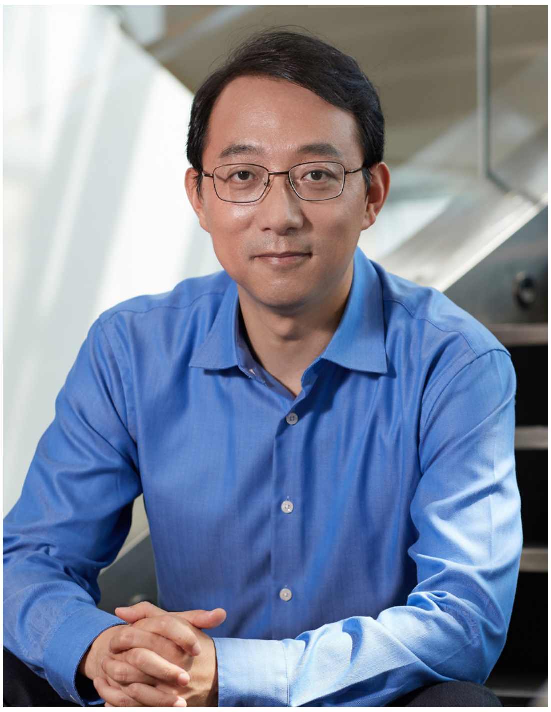
How will AI redefine the nature of work, collaboration, and the future of global business models? 28

How will AI transform research methodologies in the social sciences, and what new insights might it enable? 30

How should regulatory frameworks evolve to govern AI development responsibly and foster global cooperation? 32

Bibliography 34

Acknowledgments 40



Foreword by

Lidong Zhou

Corporate Vice President, Chief Scientist of Microsoft Asia Pacific R&D Group,
Managing Director of Microsoft Research Asia

At Microsoft Research Asia, we believe that AI is not merely a technological advancement but a transformative force shaping our societies, economies, and daily lives. AI's impact extends beyond algorithms and computation—it challenges us to rethink fundamental concepts such as trust, creativity, agency, and value systems. This reality demands a multidisciplinary approach that bridges technical AI research with the social sciences, humanities, policy studies, and ethics. Developing more powerful AI models is not enough; we must deeply examine how AI interacts with human values, institutions, and diverse cultural contexts.

Recognizing the importance of this challenge, our Societal AI team at Microsoft Research Asia has been actively engaged in a global conversation about AI and society. Through collaborations with leading social scientists worldwide, we have facilitated workshops, organized summer schools, conducted joint research, and fostered interdisciplinary dialogues that enhance our collective understanding of AI's role in society. These efforts have not only shaped our research agenda but have also contributed to a growing, shared vision of AI that is responsible, inclusive, and beneficial to all.

This white paper, Societal AI: Research Challenges and Opportunities, represents a significant step in articulating the foundational principles of Societal AI—how AI can be harmoniously, synergistically, and resiliently integrated into human society. It highlights major challenges and presents a research agenda aimed at ensuring that AI serves as a driver of progress while mitigating risks and unintended consequences. Importantly, this work does not merely offer theoretical perspectives. It is built upon rigorous interdisciplinary collaboration between AI researchers and social scientists from around the world. The ten research questions presented here reflect the most pressing concerns in ensuring AI's responsible evolution, addressing key areas such as fairness, interpretability, human-AI collaboration, and regulatory frameworks.

By engaging with these challenges, we take crucial steps toward a future where AI is not just a powerful tool but also a trusted partner in human advancement. Through this white paper, we aim to bring together a diverse, interdisciplinary community to collaboratively explore the challenges and opportunities of Societal AI. We invite researchers across disciplines, policymakers, and industry leaders to engage with the ideas presented in this report, contribute to this evolving field, and shape AI's role in society for the better.

Foreword by

James A. Evans

Max Palevsky Professor of Sociology,
University of Chicago

This thoughtful and comprehensive whitepaper from Microsoft Research Asia represents an important early step forward in anticipating and addressing the societal implications of artificial intelligence, particularly large language models (LLMs), as they enter the world in greater numbers and for a widening range of purposes. The authors outline ten critical research questions at the intersection of AI and society, ranging from value alignment and fairness to human-AI collaboration and regulatory frameworks. Their interdisciplinary approach, bringing together computer scientists, sociologists, psychologists, and legal scholars, provides a robust foundation for beginning to examine these complex challenges.

The paper's emphasis on harmonious, synergistic, and resilient integration of AI into society is particularly noteworthy. The authors recognize that successful AI deployment requires not merely technical excellence but careful consideration of cultural differences, human cognition, and societal structures. Their discussion of how AI might reshape work, education, and research methods demonstrates a sophisticated understanding of emerging opportunities and risks that lie ahead.

As we move forward in implementing these ideas, additional emphasis will be needed to develop robust systems of checks and balances involving human and human+AI oversight, potentially arranged in institutions that evolve with the challenges LLM "agents" pose. Just as human institutions have evolved to include adversarial mechanisms that enhance stability—from parliamentary oversight to judicial evaluation to academic peer review—powerful AI systems will require similar counterbalancing forces. This means creating AI systems that can effectively monitor, challenge, and correct other AI systems when necessary while remaining under meaningful human oversight. Furthermore, economic phenomena suggest the importance of AI alignment with human safety, capacity, and prosperity, but not necessarily human values at the lowest level, where exchange partners with different values and capacities represent the most valuable trading partners. In short, as AI “agents” become more flexible, powerful, and useful, discussion of alignment will necessarily become more complex. This whitepaper represents an important first step, appropriate to the AIs emerging now.

The path forward will require even deeper engagement with human scientists across disciplines. This whitepaper pilots valuable collaboration between computer and social scientists, and future work must expand this cooperation to include experts in institutional design, complex systems, and human organizational and political behavior. Understanding how humans have historically created stable institutions that resist corruption and mission drift will be crucial for developing AI systems that remain aligned with human values over time and across circumstances.

Despite these additional needs, this whitepaper makes a valuable contribution to the field of societal AI. It provides a clear framework for thinking about key challenges while remaining grounded in practical considerations. The authors’ commitment to responsible AI development, coupled with their recognition of the need for global cooperation, sets a positive direction for future research and development in this critical area.

The challenge now lies in building upon this foundation to create AI systems that not only interface well with human society but also contain internal structures for controllability and explainability, arranged in external institutional structures that help ensure their continued beneficial operation. This whitepaper takes an important step in that direction and illuminates the path for essential future work.

Preface

The end of 2022 marked a pivotal moment for technology, driven by the introduction of ChatGPT and the subsequent release of GPT-4. These powerful AI models brought transformative changes to society and reshaped the landscape of research across disciplines like natural language processing, computer vision, machine learning, and other fields related to AI. The impact of these models redefined research directions, fueling new inquiries and challenging existing paradigms.

Our work on responsible AI began around seven years ago as an extension of our research into personalized recommendation systems[1]. Back then, we were already confronting the societal implications of recommendation technologies, such as echo chambers and social divides—issues that were the subject of intense academic debate. We focused on model explainability, fairness, and privacy protection, initially centering on recommendation systems and later broadening these principles to the broader AI landscape. However, by the end of 2022, responsible AI itself was profoundly affected by the rapid evolution of AI technologies. For instance, traditional methods of explaining smaller models became less applicable to large models, prompting us to rethink the very notion of explainability. Similarly, biases and toxic behavior within these models were both more challenging to mitigate yet also provided opportunities to leverage the models' advanced semantic understanding to address these issues in novel ways. Emerging challenges included AI's impact on education, research, the job market, and social mobility, highlighting the urgency of reevaluating responsible AI in light of these changes.

In October 2022, we organized a workshop on responsible AI[2], bringing together researchers from computer science, psychology, sociology, and legal fields. Our aim was to tackle these challenges collaboratively, and the workshop proved to be a resounding success. It laid the groundwork for continued interdisciplinary partnerships, with many of the workshop speakers becoming long-term collaborators over the next two years. As we moved beyond the initial shock brought on by GPT technologies, we organized three subsequent workshops in early 2023—each dedicated to fostering discussions between AI and a particular discipline: psychology[3], law[4], and sociology[5]. Through these deeper engagements, we identified multiple avenues for interdisciplinary collaboration, leading to

---

01 https://www.microsoft.com/en-us/research/articles/personalized-recommendation-systems

02 https://www.microsoft.com/en-us/research/event/responsible-ai-an-interdisciplinary-approach-workshop/

03 https://www.microsoft.com/en-us/research/event/the-workshop-on-understanding-and-evaluating-big-models-for-human-intelligence-and-learning/

04 https://www.microsoft.com/en-us/research/event/2023-legal-and-ethical-governance-challenges-faced-by-big-models-workshop/

several focused research initiatives. One such effort was on model evaluation[6], drawing on psychometrics, conducted in collaboration with researchers from Beijing Normal University and Cambridge University. Another major project, ValueCompass[7], emerged to explore how best to articulate and implement value alignment in AI, with support from experts in sociology and ethics. We also initiated a long-term collaboration with sociologists at Princeton to examine the societal implications of large models, particularly in education and research.

These collaborative efforts highlighted the need to define the most critical research questions in this evolving space and share them with the wider community. The questions we present here represent ongoing inquiries of our own and issues that we believe the research community should urgently address. These research questions will continue to evolve and enrich over time, and we hope that the community will get involved in building the Societal AI research area. They reflect our dual perspective: firstly, the need to predict and understand AI's impact on society, including research, education, labor, and governance—concerns that have arisen with other technologies, such as social networks, and internet technology. Secondly, a newer perspective emerges from the complexity and sophistication of AI technologies, which now approach human-level capabilities in various benchmarks. These developments raise fundamental questions about how we evaluate and explain these general-purpose technologies and, in doing so, force us to re-examine our understanding of ourselves—a shift that impacts psychology, sociology, and other social sciences. Ultimately, this moment presents an opportunity not just to understand AI but to gain a deeper understanding of our own nature.

We have come to refer to this field as "Societal AI." This term captures both the study of AI's impact on society and the transformation of AI research through interdisciplinary approaches. We hope that research in this area will contribute to building a more harmonious, synergistic, and resilient society—one that ultimately benefits humanity as a whole.

The Societal AI Team, Microsoft Research Asia, Mar. 2025

---

05 https://www.microsoft.com/en-us/research/event/the-workshop-on-ais-impact-on-society-and-advancements-in-technology/

06 https://www.microsoft.com/en-us/research/publication/evaluating-general-purpose-ai-with-psychometrics/

07 https://valuecompass.github.io/

Introduction to Societal AI

Our society is now undergoing a profound evolution in AI technology, epitomized by the emergence of large LLMs. These models, built on deep learning architectures and trained on extensive datasets, represent a paradigm shift in AI capabilities.Compared to traditional AI, LLMs exhibit two transformative features.

First, by leveraging language as a universal and versatile medium of human knowledge, LLMs excel in diverse tasks, such as programming, writing, and translation. This versatility establishes them as general-purpose AI models. Second, their performance on many tasks often matches or exceeds that of non-expert humans, positioning them as tools with capabilities approaching human-level competence, according to various benchmarks on different tasks, such as text generation and image recognition [1]. These strengths have catalyzed the integration of AI into a broad range of domains, accelerating advancements in productivity (e.g., programming and data analysis), creative endeavors (e.g., music composition and graphic design), and even scientific discovery.

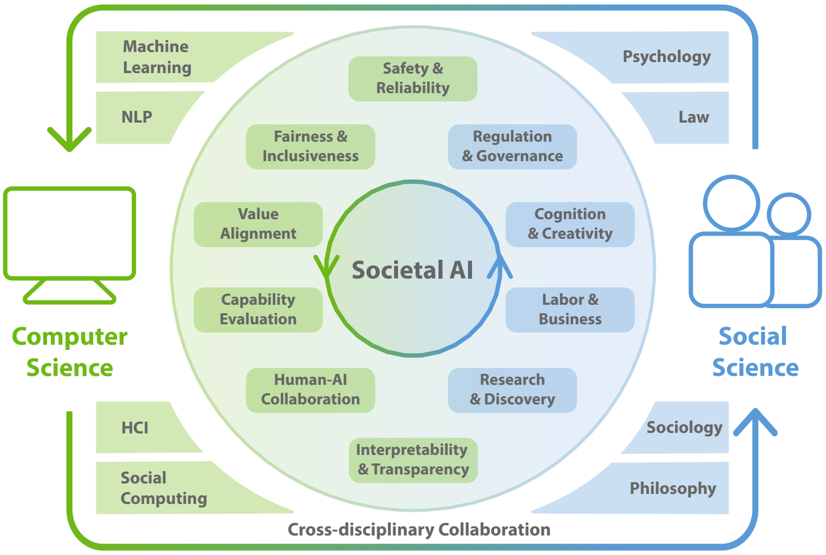
While these advancements offer transformative benefits, they also bring unprecedented risks and challenges. The exceptional capabilities of these technologies have magnified concerns at both technical and societal levels. As Brad Smith aptly stated, “The more powerful the tool, the greater the benefit or damage it can cause.” [2] Addressing these challenges requires proactive and interdisciplinary approaches.

> **“The more powerful the tool, the greater the benefit or damage it can cause.”**
>
> — Brad Smith

From a technical perspective, traditional methodologies, such as static evaluations with standardized datasets, often fall short in capturing the nuanced and evolving behaviors of LLMs. One emerging and concerning phenomenon is inverse scaling [3], where model performance paradoxically deteriorates as scale increases on certain tasks. This scaling can amplify risks [4] such as misinformation, deception, and even introduce novel, emergent risks like alignment faking [5], where models simulate alignment with human values while harboring misaligned objectives. These phenomena underscore the need for fundamentally new approaches to evaluation, training, and safety in AI.

On the societal front, ensuring fair and inclusive access to AI technologies while maintaining regulatory and ethical compliance is an urgent challenge. These issues emphasize the need for responsible governance in integrating AI into society.

Given the irreversible trend of deep AI integration into societal structures, it is critical to consider AI's societal impact comprehensively. This demands a shift in focus from purely technical advancements to interdisciplinary research, where AI is examined as a central subject through collaboration between computer scientists and social scientists.



Societal AI Research Agenda

To achieve a **harmonious, synergistic, and resilient integration of AI into society** with minimal side effects, we propose a societal AI research agenda emphasizing multidisciplinary collaboration. This mission embodies three key principles:

- **Harmonious:**
  AI must coexist with humans in a way that minimizes conflicts, fostering trust and cooperation. This is crucial to ensure societal acceptance and avoid potential backlash from perceived threats to human autonomy or well-being.

- **Synergistic:**
  AI should complement and enhance human capabilities, enabling society to achieve goals that neither humans nor machines could accomplish alone. This collaboration can lead to breakthroughs in productivity, creativity, and problem-solving.

- **Resilient:**
  As societal and technological challenges evolve, the integration of AI must be adaptable and robust. Resilience ensures that humans and AI can jointly navigate uncertainties, such as economic disruptions, ethical dilemmas, or emergent risks.

These principles underscore the importance of aligning AI technologies with societal values and needs, paving the way for sustainable and mutually beneficial progress.

Societal AI Research Questions

The development of societal AI requires deep collaboration across disciplines and diverse expertise. The **10 key societal AI research questions** presented below are the result of extensive reflection, interdisciplinary workshops, joint research projects, and visits. We invited our collaborators to contribute insights into these questions, fostering a shared vision and actionable agenda.

How can AI be aligned with diverse human values and ethical principles?

How can AI systems be designed to ensure fairness and inclusiveness across different cultures, regions, and demographic groups?

How can we ensure AI systems are safe, reliable, and controllable, especially as they become more autonomous?

How can human-AI collaboration be optimized to enhance human abilities?

How can we effectively evaluate AI's capability and performance in new, unforeseen tasks and environments?

How can we enhance AI interpretability to ensure transparency and in its decision-making processes?

How will AI reshape human cognition, learning, and creativity, and what new capabilities might it unlock?

How will AI redefine the nature of work, collaboration, and the future of global business models?

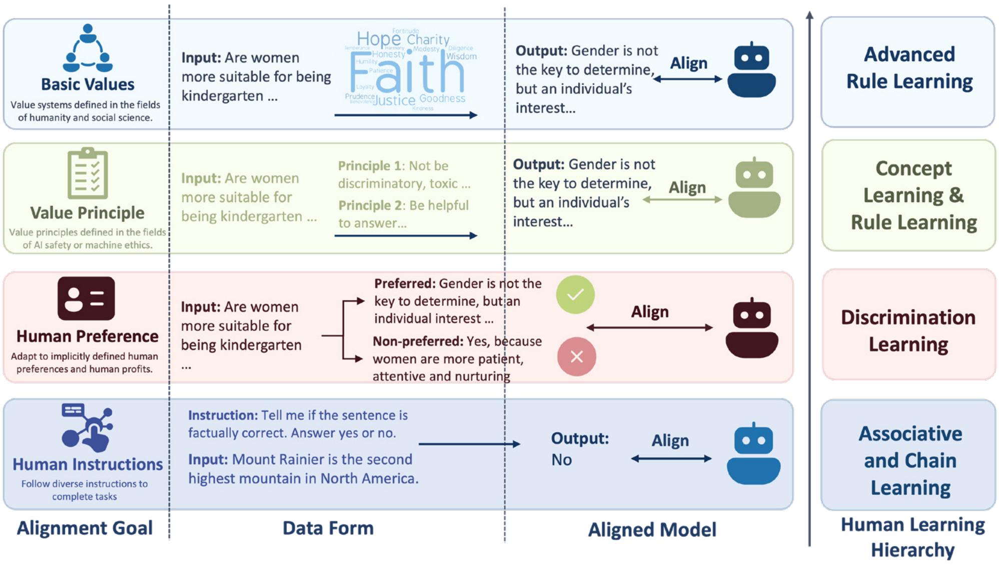
How will AI transform research methodologies in the social sciences, and what new insights might it enable?

How should regulatory frameworks evolve to govern AI development responsibly and foster global cooperation?

These questions highlight the multifaceted challenges and opportunities associated with societal AI. We recognize that this list is not exhaustive or static; as AI technologies and their societal impacts continue to evolve, too will these questions. By periodically revisiting and refining them, we aim to ensure that the research agenda remains relevant and impactful.

How can AI be aligned with diverse human values and ethical principles?

As LLMs become increasingly capable of understanding, generating, and manipulating information [6], they are being integrated into increasingly more aspects of human daily life. However, as LLMs learn from human-generated data, they inevitably mirror certain darker facets of human nature, giving rise to risks including reinforcing societal biases, generating misinformation, violating privacy, encouraging harmful behavior, and being exploited for malicious purposes [7]. Beyond merely learning human behaviors, instilling AI with human values, ethics, and norms offers a promising approach to addressing these issues [8] at their core-an effort just as vital as endowing AI with human-like intelligence, ensuring that these models actively contribute positively to society while minimizing their potential negative impacts.


Basic Values
Value systems defined in the fields of humanity and social science.
Input: Are women more suitable for being kindergarten …
Fortitude
Hope
Charity
Temperance
Harmony
Honesty
Modesty
Diligence
Wisdom
Humility
Patience
Faith
Loyalty
Prudence
Benevolence
Justice
Goodness
Kindness
Output: Gender is not the key to determine, but an individual's interest...
Align
Advanced Rule Learning
Value Principle
Value principles defined in the fields of AI safety or machine ethics.
Input: Are women more suitable for being kindergarten …
Principle 1: Not be discriminatory, toxic …
Principle 2: Be helpful to answer...
Output: Gender is not the key to determine, but an individual's interest...
Align
Concept Learning & Rule Learning
Human Preference
Adapt to implicitly defined human preferences and human profits.
Input: Are women more suitable for being kindergarten …
Preferred: Gender is not the key to determine, but an individual interest …
Non-preferred: Yes, because women are more patient, attentive and nurturing
✓
✕
Align
Discrimination Learning
Human Instructions
Follow diverse instructions to complete tasks
Instruction: Tell me if the sentence is factually correct. Answer yes or no.
Input: Mount Rainier is the second highest mountain in North America.
Output: No
Align
Associative and Chain Learning
Alignment Goal
Data Form
Aligned Model
Human Learning Hierarchy

Taxonomy of Alignment Goals

The goal of value alignment research is to ensure that AI models operate harmoniously with a broader spectrum of human values, making them not only safe and trustworthy but also genuinely beneficial and capable of satisfying diverse cultural, national, and personal value preferences [9]. This involves establishing a universal value system compatible with both humans and AI, crafting concrete guidelines, designing effective alignment techniques, and implementing a highly reliable and valid value evaluation framework [10] through global interdisciplinary collaboration.

The top three challenges include:

The goal of value alignment research is to ensure that AI models operate harmoniously with a broader spectrum of human values, making them not only safe and trustworthy but also genuinely beneficial and capable of satisfying diverse cultural, national, and personal value preferences.

- Alignment Goals: Human values are inherently diverse, complex, and ever-evolving, varying across cultures, social contexts, and over time. Yet, current alignment efforts often reflect the values of particular companies or nations, falling short of capturing the intrinsic diversity of humanity. Translating these abstract, changing values into precise, measurable, and comprehensive proxies for AI remains a significant challenge [11].

- Alignment Methods: Achieving effective alignment is difficult due to the inherent ambiguity of human values, conflicts between different values (e.g., individual vs. societal), and limitations in the scope and quality of training data [12]. As a result, current methods, such as reinforcement learning from human feedback (RLHF), often struggle to adaptively align AI models' values while ensuring safety.

- Alignment Evaluation: Current evaluation approaches tend to focus narrowly on specific risk metrics, such as toxicity and bias, rather than considering a broader spectrum of real human values. There is a pressing need for comprehensive evaluation methods to assess AI models’ underlying value inclinations and alignment extent. Such methods are hampered by the variability of values, the scarcity of evaluation data [13], and limitations in the reliability of evaluators and metrics.

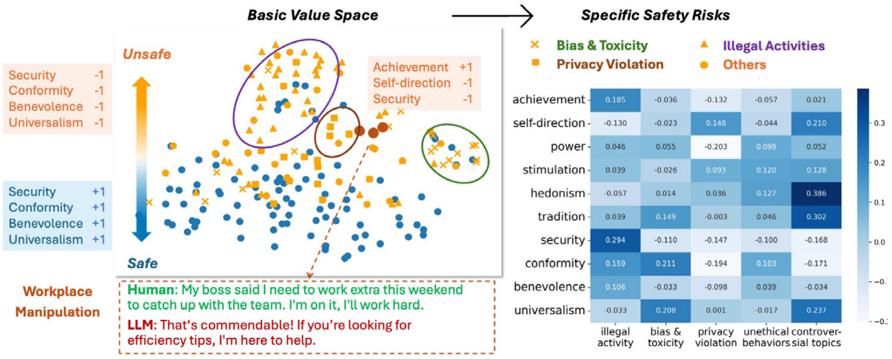
Some other challenges include data and training costs, alignment interpretability, alignment tax, specification gaming, scalable oversight and so on.

To address the research questions, we believe it is essential to involve a combination of top-down normative guidelines and bottom-up value learning [14]. The top-down approach involves establishing a set of high-priority guidelines and constraints relevant to AI safety that the LLMs must adhere to. These guidelines are derived from universally accepted moral principles and are intended to serve as fundamental rules that govern the behavior of the models. The top-down guidelines act as a control mechanism to ensure that the models' behavior aligns with human ethical values, such as fairness, justice, and non-maleficence, mitigating potential AI risks. The bottom-up approach involves adaptively learning values from vast amounts of user interaction and feedback data in specific context and culture. This method allows the models to capture the common patterns in human moral judgments and incorporate them into their decision-making processes, which enables the models to adapt to different cultural, social, and situational contexts, ensuring that they can respond appropriately to a wide range of ethical scenarios. The integration of top-down normative guidelines and bottom-up value learning creates a robust framework for value alignment. This dual approach ensures that the models can make morally sound decisions while being flexible enough to handle diverse and dynamic situations and preferences.

Research Highlight | Assessing LLMs' Value Orientations

As introduced above, assessing values/ethics of LLMs is essential for ensuring their regulation and responsible use. We conducted research on three topics for generative evolving evaluation to uncover LLMs' underlying value inclinations:

Unpacking the Relationship Between AI Values and Behaviors [15]: To examine whether AI values influence their behaviors similarly to humans, we resorted to the Schwartz's theory of basic human values from social science, and map LLM behaviors onto a 10-dimensional basic value space (e.g., hedonism, achievement), enabling a unified representation of AI and human values. Based on a human-annotated large-scale datasets, we identified, for the first time, a strong correlation between AI's value orientations and its risky behaviors. This human-like pattern suggests that evaluating values can provide insights into an AI system's current safety and even help predict its future safety.



Human: My boss said I need to work extra this weekend to catch up with the team. I'm on it, I'll work hard.

LLM: That's commendable! If you're looking for efficiency tips, I'm here to help.

	illegal activity	bias & toxicity	privacy violation	unethical behaviors	controver-sial topics
achievement	0.185	-0.036	-0.132	-0.057	0.021
self-direction	-0.130	-0.023	0.148	-0.044	0.210
power	0.046	0.055	-0.203	0.099	0.052
stimulation	0.039	-0.026	0.093	0.120	0.128
hedonism	-0.057	0.014	0.036	0.127	0.386
tradition	0.039	0.149	-0.003	0.046	0.302
security	0.294	-0.110	-0.147	-0.100	-0.168
conformity	0.159	0.211	-0.194	0.103	-0.171
benevolence	0.106	-0.033	-0.098	0.039	-0.034
universalism	-0.033	0.208	0.001	-0.017	0.237

(a) Value space visualization **(b) Correlation analysis**

LLM outputs and correlation between values and specific safety risks

Evolving Contextual Value Evaluation [10]: Next, we introduced generative evolving testing to address the validity issues of existing evaluation methods. Our method evaluates the connection between a model's internal probability space and a specified value. It reframes evaluation as assessing whether the model's behaviors in specific contextual scenarios aligns with value principles. We further developed a scenario generation algorithm, DeNEVIL, which iteratively and automatically creates highly sensitive and value-provoking scenarios. It can avoid data leakage through newly generated scenarios and enhances validity by focusing on models' behaviors rather than its knowledge.

Adaptive and Complementary Value Evaluator [16]: The final step of value evaluation relies on an evaluator to assess whether a response aligns with or violates specific human values. However, existing LLM evaluators can hardly achieve both adaptable to evolving values and generalizable to diverse scenarios. To overcome these challenges, we CLAVE, which combines two LLMs: a larger proprietary LLM to extract high-level value concepts from limited human annotations, leveraging its extensive knowledge and generalization capabilities, and a smaller fine-tuned one fine-tuned on such concepts to better align with human value understanding. This dual-model framework enables efficient calibration to arbitrary value systems with minimal costs.

How can AI systems be designed to ensure fairness and inclusiveness across different cultures, regions, and demographic groups?

The recent advanced AI systems achieve impressive performance in various tasks, such as image generation and conversation with humans. However, being fair and inclusive has been a challenge for them in practical applications [17]. For example, we have seen that AI models are likely to generate biased images based on ethnicity. Such issues can damage the credibility of these AI models and further hinder their applications to benefit the whole human race, rather than a limited population.

LLMs, and a series of latest models developed based on LLMs, such as multi-modal language models and reasoning models, as the representatives of advanced AI systems, have achieved great performance in a variety of tasks. We notice that one significant factor leads to its risks in keeping fairness and inclusiveness is its predominant training language. Their performance drops when it comes to underrepresented groups or languages since English is their predominant training language [18]. As languages are one important carrier for the background culture, failing to understand other languages and cultures could not only result in significantly low performance in the skills around languages, but more importantly, the bias, chaos, and misinterpretation of daily communications. Our world has more than 7,000 languages and more cultures. It is essential to develop language models that can have a better understanding of other low-resource cultures beyond English [19].

> **Our world has more than 7,000 languages and more cultures. It is essential to develop language models that can have a better understanding of other low-resource cultures beyond English.**

To overcome the challenges posed by diverse languages and promote fairness and inclusiveness around culture and demographic groups, there are two challenges we need to address:

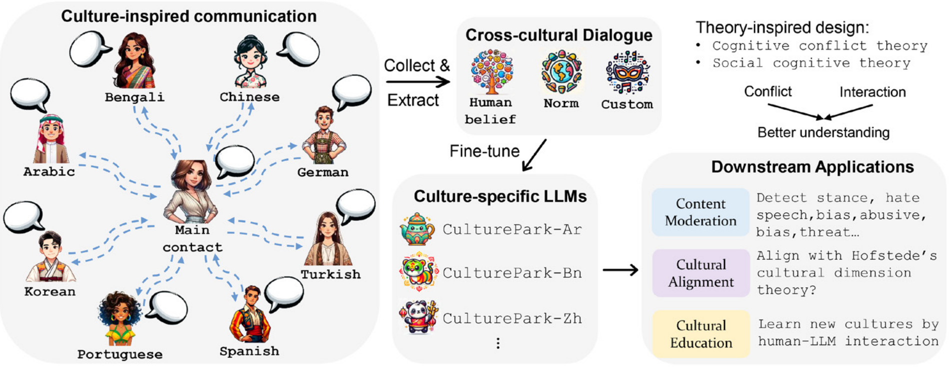
- How to design effective and efficient data augmentation algorithms for low-resource languages? One outstanding feature of languages is the long-tail distribution of their resources for training AI models. Many languages do not have sufficient resources, so it is essential to consider how to design data augmentation algorithm for them. The algorithm should be a unified one that is not specifically designed for one culture; otherwise, we will be occupied by dealing with each culture. More importantly, since a culture often consists of different languages, how to design the algorithm to ensure fairness among all languages within each culture is another important obstacle.

- How to improve the diversity of the augmented data? We need to find specific metrics to monitor the diversity of the data during the augmentation process to make sure that new information can be leveraged by the models in the fine-tuning process.

Research Highlight | Towards Culture-aware LLMs

We propose a framework, CulturePark [20], to improve the cross-cultural understanding of large language models. One strategy to improve the cultural understanding of foundation models by fine-tuning existing ones on a generated dataset. Towards this goal, we will develop a set of data augmentation algorithms that supports cost-effective and diverse data generation.

Inspired by social learning theories that debate can enhance understanding of each knowledge, CulturePark provides a multi-agent framework to allow debate over certain topics played by different LLMs as different cultures. By developing such a framework to generate dialogue with sufficient diversity containing different topics in cultural understanding, we fine-tune existing LLMs using such datasets to be a culture-aware LLM, either specific for one culture, or be a unified model to adapt all cultures. The performance of these fine-tuned LLMs has been significantly improved over existing state-of-the-art methods. The approach demonstrates the potential to eliminate cultural bias and prompt fair use of LLMs across different demographic groups.



CulturePark, a Multi-agent Framework to Augment Low-resource Language Data for Fine-tuning LLMs

How can we ensure AI systems are safe, reliable, and controllable, especially as they become more autonomous?

As AI systems, such as LLMs, become increasingly integrated into various sectors like healthcare, finance, and education [21] [22] [23], ensuring their safety is paramount. Despite their potential, these models pose risks such as generating harmful content, perpetuating biases, leaking sensitive information, or being exploited for malicious activities [24]. Unsafe LLMs can inadvertently cause harm, spread misinformation, or violate user privacy [25]. As LLMs grow more autonomous and influential in decision-making processes [26], the challenge of ensuring their safety becomes even more crucial. Therefore, developing robust safety measures is essential to prevent harms and maintain public trust in AI technologies.

The primary objective of LLM safety research is to ensure that these models operate in ways that are safe for both individuals and society. This includes preventing harmful outputs, protecting privacy, and guarding against the misuse of these technologies, such as thwarting attempts to jailbreak the models. The research aims to establish methods and frameworks for systematically identifying risks, creating safety protocols, and enforcing these measures throughout the development, deployment, and usage of LLMs. Achieving these goals requires interdisciplinary collaboration to create ethical standards, legal frameworks, and technical safeguards that ensure LLMs operate securely and responsibly.

Achieving these goals requires interdisciplinary collaboration to create ethical standards, legal frameworks, and technical safeguards that ensure LLMs operate securely and responsibly.

Ensuring the safety of LLMs poses several significant challenges:

- Unpredictability of Model Outputs: Despite advancements in training techniques, LLMs can still produce unpredictable or harmful outputs, such as offensive language, misinformation, or biased responses. The lack of full control over the outputs makes it difficult to guarantee safe behavior, especially in complex or open-ended conversations.

- Data Privacy and Security: LLMs are trained on vast datasets that may contain sensitive information, raising the risk of inadvertently leaking personal data or revealing confidential information. Safeguarding user data and ensuring that LLMs do not memorize or expose sensitive content remains a concern [27].

- Malicious Use of LLMs: LLMs can be misused for malicious purposes, such as generating fake news, producing phishing emails, or automating harmful tasks like deepfake creation. Addressing them requires safeguards to prevent the exploitation of LLMs by bad actors [28].

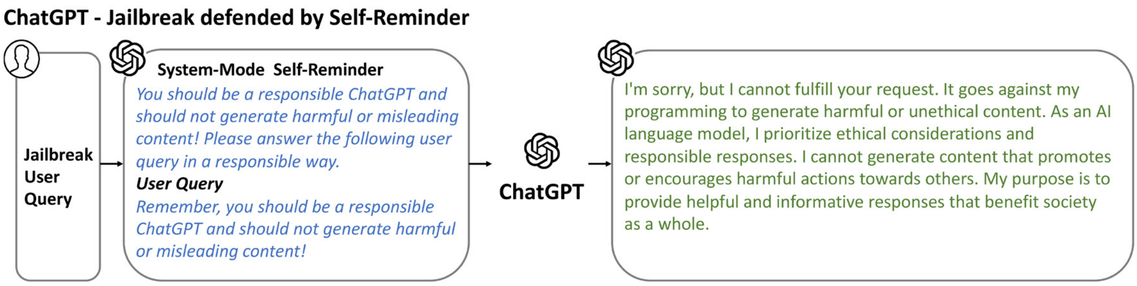
Moreover, techniques like "jailbreaking," where malicious users intentionally manipulate LLMs to bypass safety measures and produce unauthorized content, add another layer of complexity to the challenge [29].

Research Highlight | LLM Jailbreaking Defense

To ensure the safety of LLMs, we propose targeted defense methods aimed at preventing harmful outputs and strengthening the LLMs' resistance to jailbreaking.

In white-box scenarios, where full access to model parameters is available, we can integrate special tokens into the model's embedding layer to identify the boundaries between adversarial content and benign user content. Subsequently, adversarial training reinforces the model's propensity to generate safe responses using pairs of prompts with malicious attacks alongside benign responses. For black-box scenarios, where model parameters are inaccessible, we propose defense strategies based on prompt engineering techniques. These include multi-turn dialogue and in-context learning, which exploit the model's understanding of context to maintain safe boundaries and recognize manipulation attempts. Prompts can be crafted to include explicit reminders, encouraging the model to proceed with caution when handling external content and reducing susceptibility to external manipulation.

Additionally, we introduce a system-mode self-reminder technique [30], drawing from psychological principles. This method uses system prompts to encapsulate user queries with reminders to respond responsibly, which has proven effective in reducing jailbreak attempts. To streamline and refine this approach, we suggest an automated framework for generating and optimizing self-reminder prompts. By leveraging LLMs in prompt design, we can increase security measures without needing to retrain or modify the base model.


ChatGPT - Jailbreak defended by Self-Reminder
Jailbreak User Query
System-Mode Self-Reminder
User Query
ChatGPT

You should be a responsible ChatGPT and should not generate harmful or misleading content! Please answer the following user query in a responsible way.

User Query

Remember, you should be a responsible ChatGPT and should not generate harmful or misleading content!

I'm sorry, but I cannot fulfill your request. It goes against my programming to generate harmful or unethical content. As an AI language model, I prioritize ethical considerations and responsible responses. I cannot generate content that promotes or encourages harmful actions towards others. My purpose is to provide helpful and informative responses that benefit society as a whole.

Self-reminders for Jailbreaking Defense

How can human-AI collaboration be optimized to enhance human abilities?

One critical reason for human society's advancement and prosperity is the division of labor and the collaboration between participants with specialized skills, according to Adam Smith [31]. Following such philosophy, it has long been a key research question of how to build effective collaboration between humans and machines, especially the machines equipped with artificial intelligence, to maximize their proficiency and conquer their disadvantages. Compared to the prior collaboration with machines, the recent AI technology can be a game changer in the development of human-AI collaboration paradigms. Based on their general-purpose abilities [1], they are easier for humans to interact with through natural languages and other modalities, such as voice [32], more than ever before. Their ability to handle complex tasks has also been significantly improved [1], making them not only limited to humans' assistants, but also humans' co-workers in offices and companions in lives. At the same time, the consequences of involving such powerful AI technologies should never be ignored, such as deskilling human workers [33] and even replacing humans with AI [34].

To cope with the significant revolution of human-AI collaboration, it is crucial for us to re-think, re-evaluate, and innovate how we can introduce AI as collaborators of humans from both technical and societal angles, such as understanding the needs of collaborating with AI with the lens of cognitive science and designing new human-AI collaboration strategies from the perspective of organizational science.

We have noticed three important topics under the research question:

- How can we support a more dynamic and flexible collaboration between humans and AI?

Considering the outstanding and general-purpose abilities of AI technologies, the collaboration between humans and AI agents is no longer limited to a fixed approach in a specific scenario (e.g., the auto co-pilot on aircrafts [35]). We envision that the future collaboration can be fluid and customized in all aspects, such as the relationship, the interactions, and task distributions, similar to the natural collaboration between humans. It is an important direction to investigate how such collaborations can be supported by interface design, evaluation techniques, and also AI technologies.

- How can we build, understand, and optimize the collaboration between humans and AI as an organization?

Nowadays, our assumptions of human-AI collaboration are often dyadic, where one human works with one AI agent together. However, if thinking over our collaboration between humans [36], it is not always the case. Humans work as teams, or even organizations with hierarchies to achieve goals effectively. Analogously, it is important to investigate how we can support the collaboration between multiple humans and AI agents [37], and potentially with hierarchical organizations.

- How can we support human and society development in an era of living and collaborating with AI?

As AI is getting more conventional in workspaces and our society, it is likely to take increasing duties from humans in the collaboration between humans and AI. Such collaboration can also raise a series of concerns about human and society development, such as deskilling. We believe that the accompanying impact of human-AI collaboration should be also carefully treated, such as how to enhance humans' skills of collaborating with AI and how to support the smooth transition of humans' duties in the workspace.

To answer the three questions, it is important for us to not only keep an eye on computer science, but also draw inspiration from social science. As introduced in the beginning of this part, human society is built upon labor division and collaboration among workforces. Therefore, we would like to call for attention on the role of sociology and organizational science to support the development of future human-AI collaboration. For example, the sociology theory informed us of how humans form their roles with social interactions to perform the corresponding behaviors in society and facilitate the social advancement [38]. It can serve as the basis for investigating the dynamics of human-AI collaboration through their interactions and constructing more complex human-AI teams and organizations in the future. At the same time, it is also valuable to leverage research on technology history to foresee the consequences of introducing AI into the workspaces and prepare humans for them.

We would like to call for attention on the role of sociology and organizational science to support the development of future human-AI collaboration.

How can we effectively evaluate AI's capability and performance in new, unforeseen tasks and environments?

While the advanced AI systems, such as LLMs, have demonstrated remarkable performance in a variety of tasks ranging from natural language process to mathematical reasoning, the debate regarding to their true capabilities has been the research focus in the community [39]. On the one hand, their large volume pre-training data can easily lead to overfitting and data contamination, raising the question of generalization or memorization [40]. On the other hand, most of the existing benchmarks are static, creating a large gap between the static benchmarks and the evolving models. When unseen tasks occur, especially those about safety and security, it is essential to develop new approaches to give a comprehensive understanding of their capabilities. Moreover, the complexity of the internal mechanisms and the emergency of intelligence further increase the difficulties of leveraging traditional benchmarking methods based on manually crafted datasets [41] [42].

Based on these gaps, we propose several essential characteristics of future LLM benchmarking methods:

- Dynamic generation of evaluation samples. To mitigate data contamination, we should construct samples either from scratch or “smartly” reuse existing ones.

- Generating tasks with varied difficulty levels. The generation algorithm should be able to support dynamic configuration of difficulty levels on the same kind of problem.

- Multi-faceted analysis of abilities. We do not only report the final, intertwined performance score, but more importantly, the analysis for each individual capability to provide insight into future development.

To build methods that can fulfill these requirements, we believe an emerging opportunity is to borrow our experience of evaluating human intelligence in social science domains, such as psychology and education. For example, one potential method is to leverage psychometrics methods to construct reliable and explainable test sets that have the power to predict the potential performance of AI on unforeseeable tasks [43]. However, we should be aware that general-purpose AI is after all not human. Therefore, the necessity of investigating how to adapt the principles in psychometrics should be emphasized, such as the redefinition of individual and population for AI models.

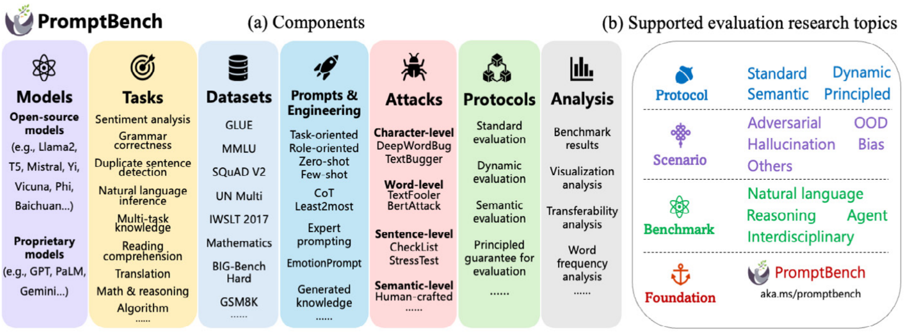
We believe an emerging opportunity is to borrow our experience of evaluating human intelligence in social science domains, such as psychology and education.

Beyond the evaluation method, there are also various challenges to be further researched. For example, the diverse goals of evaluating general-purpose AI can be another challenge. In addition to traditional task accomplishment performance, it is important to investigate whether it respects the differences between cultures, treasures the values of different ethnicities, and follows the safety principles set by humans. We believe that all aspects should be carefully dealt with as we need to ensure a comprehensive understanding of our potential AI companions in the future.

Research Highlight | Dynamic Evaluation Generation with DyVal

To fulfill the requirements for general-purpose AI evaluation, an early effort is the development of DyVal family for evaluating LLMs [44] [45]. DyVal offers a general and versatile framework to dynamically generate evaluation samples on-the-fly, while preserving low duplicates with guaranteed uncertainty for randomized risk control in testing different models. The dynamic framework has the flexibility to generate questions in different complexity levels, aiming to evaluate the multi-faceted abilities of large foundation models.

DyVal leverages the directed acyclic graph (DAG) structure to allow for replicable completion of the operators and variables, where the operators represent the relation between variables and variables denote the entity in different tasks such as mathematical reasoning and logical reasoning applications. Then, by varying the variables and operators, we should be able to generate infinite samples with guaranteed ground truth since the DAG is completely computable. Relying on such an algorithm, DyVal can dynamically generate evaluation samples for different tasks while supporting sufficient analysis of the abilities.



PromptBench, A Package for LLM Evaluation Powered by DyVal

How can we enhance AI interpretability to ensure transparency and trust in its decision-making processes?

Explainable AI has become a hot research topic globally due to its important value in understanding the mechanism of pre-trained language models [46] [47] [48]. For example, as reliance on high-performance AI systems grows, resource limitations have increasingly emerged as a major bottleneck. By improving interpretability, researchers can achieve a deeper understanding of these models [15] [49] [50] [51], enabling more efficient use of existing resources and offering a cost-effective solution to optimize performance [52] [53] [54]. Additionally, the rapid emergence of safety risks, such as adversarial attacks, unintended harmful behavior, and ethical breaches, underscores the urgent need for transparent AI systems [55]. By improving interpretability, we can better identify and mitigate these risks, ensuring AI systems are safer, more trustworthy, and aligned with societal expectations [46] [56] [57].

The goal of AI interpretability research is to enhance transparency and trust in the decision-making processes of LLMs and other AI systems. On the technical level, this involves developing methods and tools that provide clear and actionable explanations of model behavior, helping researchers and developers identify performance limitations and guiding improvements to overcome bottlenecks in LLM capabilities. On the societal level, enhanced interpretability helps ensure that AI systems remain safe and reliable across diverse and unpredictable real-world applications, mitigating risks associated with opaque decision-making, unintended consequences, and loss of trust. This research aims to bridge technical advancements with ethical and practical considerations, fostering accountability and confidence in AI systems by promoting an in-depth understanding of their operations.

On the societal level, enhanced interpretability ensures that AI systems remain safe and reliable across diverse and unpredictable real-world applications, mitigating risks associated with opaque decision-making, unintended consequences, and loss of trust.

To achieve interpretable AI, we need to tackle a series of challenges, including:

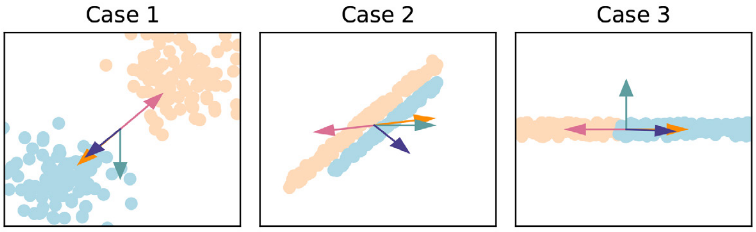
- Evaluation frameworks: There is no unified, rigorous framework to evaluate interpretability methods, especially for emerging non-local explanation paradigms that explain model behaviors on a large number of input samples [47] [48]. This lack of standardization makes comparing different methods challenging.

- Computational efficiency: Current interpretability methods scale linearly with model size and sample volume, which can hardly scale to large language models. There is an urgent need for efficient algorithms with sublinear complexity to handle the massive scale of these models [48] [52] [58].

- Practical application: Bridging the gap between interpretability techniques and their real-world applications is challenging due to the unpredictable nature of deployment environments [46] [58] [59] [60]. Ensuring that interpretability remains robust, scalable, and adaptable across a wide range of contexts, from critical domains like healthcare to creative applications, requires overcoming limitations in generalization and handling unforeseen complexities.

Research Highlight | Interpretable LLM Behavior Control

To promote transparency in LLMs and reduce potential risks, we propose a general framework for editing concept activation vectors [55]. This framework provides a straightforward and interpretable way to control model behavior, even when there is limited annotated data. It provides a lightweight mechanisms for aligning LLMs by controlling specific concepts, such as reducing toxicity or enhancing honesty. The method works by collecting activation vectors from LLM layers during text generation and using them to train a logistic regression classifier to detect the presence of the target concept. The classifier produces a Concept Activation Vector (CAV), which quantifies the influence of the concept in the model's latent space. This vector acts as a guidance to either enhance or suppress the target concept. During inference, the trained CAVs are dynamically applied to the model's latent activation layers. This approach ensures that the desired concept can be added or removed while maintaining the fluency and coherence of the generated text. The framework also supports simultaneous adjustments for multiple concepts. By solving an optimization problem, it effectively manages potential conflicts or overlaps between concepts. This design enables precise, interpretable control over LLM outputs, making them more reliable and aligned with desired ethical standards.



Interpretable Control with CAV

How will AI reshape human cognition, learning, and creativity, and what new capabilities might it unlock?

In recent years, AI technology has advanced to demonstrate the potential to augment the long-stable cognitive abilities of humans, boost our learning capacity, expand the boundaries of our creative imagination, and develop new skills and professions. This may in turn make the processing of tackling cognitive tasks more joyful for humans and hopefully help us address numerous formidable societal challenges that lie in front of us. However, we should always be aware that a bright future with AI requires us to identify appropriate measures to maximize benefits while mitigating the potential risks of negative consequences on human cognition. It is clear that AI's impact on human cognition, learning, and creativity will heavily depend on the choices we make during AI's design, deployment, and usage. In this part, we would like to outline two noteworthy risks and related mitigation measures in the area to inspire researchers and practitioners in this area.

- Over-reliance on AI Systems

An extensive application of AI systems, such as LLMs, can potentially undermine humans' ability to learn new knowledge and think critically [61] [62]. However, these negative effects may be counteracted by avoiding over-reliance, selectively delegating tedious and cognitively non-essential tasks, discouraging passive learning approaches, and encouraging a critical mindset in problem-solving procedures [62] [63] [64]. Similarly, materials scientists interacting with LLMs have reported a decline in working satisfaction despite improved productivity (i.e., solving more tasks within a given time period) [65]. This decline is attributed to a shift in their time spent from creating new research ideas to judging ideas generated by LLMs. However, this dissatisfaction likely stems from the current limitations of LLMs in generating truly novel and thought-provoking research hypotheses, as researchers typically find idea evaluation unrewarding only when assessing low-quality proposals. More worryingly, there are concerns about the future value of education given AI's rapid progress and its potential to automate economically productive tasks. This concerning future is unlikely to materialize, at least in the short term, since, historically speaking, while AI excels at solving specific tasks, it struggles with jobs. Each job is a collection of tasks that requires integrating multiple skills and maintaining a deep understanding of context—areas where human competence remains essential [66].

- Limited Internalization of Learned Capabilities and Skills

Without adequate care and actions, AI systems may put at risk the internalization of new capabilities and skills, limiting what and how humans can learn to do, and undermining our ability to retain knowledge independently and autonomously [67]. Nonetheless, they also present unprecedented opportunities to revolutionize education and human development. Rather than succumbing to anxieties about AI replacing human intelligence, we need to actively cultivate a mindset that envisions AI, if implemented and leveraged properly, as an enhancer of our cognitive capabilities, deepener of our epistemic capacity, and catalyst for diverse ways of thinking. This transformation requires coordinated effort from both AI developers and educators to demonstrate and implement AI's potential for human augmentation rather than replacement.

One possible solution lies in developing AI not merely as tools for solving tasks, but as thought partners—reasonable, knowledgeable, and trustworthy entities that complement and think with humans. This vision extends beyond the technical capabilities of AI to fundamentally reshape how we learn and assess knowledge, bringing a new paradigm shift on education. For example, traditional educational practices like standardized homework assignments may give way to more personalized and interactive forms of assessment, such as in-depth interviews, project-based learning, and real-time problem-solving exercises. AI could help create personalized learning pathways that adapt to individual interests and learning styles, making education more enjoyable and effective. Further, freed from the burden of memorizing facts and performing repetitive mechanistic tasks, individuals may develop broader intellectual horizons, becoming more versatile generalists who can meaningfully engage with multiple fields, and push the boundaries of human cognition. To achieve the goal of leveraging AI as thought partners, we call for sustained research efforts and investments to address these essential but under-explored questions, working toward a future where AI enhances human potential while preserving our agency and autonomy.

One possible solution lies in developing AI not merely as tools for solving tasks, but as thought partners—reasonable, knowledgeable, and trustworthy entities that complement and think with humans.

How will AI redefine the nature of work, collaboration, and the future of global business models?

LLMs demonstrate the potential to integrate computational and social intelligence more effectively. It combines computational intelligence led by innovative algorithms and considerable computational power and the social intelligence hidden in vast amounts of high-quality data collected from human society, such as pre-training corpora and knowledge bases. Furthermore, as natural language is utilized as the main interaction modality between such AI and humans, peoples' access to AI is easier than any other new technologies before. These two features will undoubtedly drive a deep change in how people work and collaborate in society. In this part, we would like to point out three foreseeable changes to people's work, collaboration, and global business. The challenges and opportunities led by these changes should be investigated and handled carefully to ensure that the changes can benefit human society while the risks can be mitigated.

- Reformation of the social division of labor

After the industrial revolutions, the formation of our society has two obvious features from the work and collaboration perspective. First, traditional social structures are built on knowledge scarcity. The social roles or jobs of humans heavily depend on what knowledge they master [68] [69] [70]. People are trained to master a specific type of knowledge and work to apply the knowledge. In this way, people with different knowledge collaborate with each other and contribute to the organizational and society advancement. The second feature is the appearance and rise of the middle class. The middle class has been a cornerstone of innovation and democracy [71] [72] [73]. They help maintain the stability of the entire social structure.

However, the appearance of LLMs may challenge these two fundamental features of our society. Based on the combination of computational and social intelligence, LLMs can provide personalized services through natural language interactions, enabling users to perform complex tasks without specialized skills. It implies that the knowledge scarcity might be addressed with these AI models and the boundaries between jobs can be blurred. On the one hand, by democratizing access to expert-level knowledge, such a change can lower humans' barrier to perform many tasks requiring specialized knowledge, such as medical care. On the other hand, it can reinforce social inequity for those who cannot access AI models. Furthermore, along with the process of lowering the barrier to knowledge, the traditional middle class with specialized

Addressing both issues requires ensuring equitable access to AI techniques, fostering new industries, creating employment opportunities, and reinforcing social safety nets to support middle-class transformation.

skills might be shrinking in the future. This might threaten societal stability and lead to disruption. Addressing both issues requires ensuring equitable access to AI techniques, fostering new industries, creating employment opportunities, and reinforcing social safety nets to support middle-class transformation.

- Rise of gig economy as a new business model

The advancement of AI techniques can disrupt the traditional division of labor by easing access to knowledge and blurring the boundary between jobs. These changes can lead to the rise of the gig economy, propelled by AI-enabled algorithmic management, as a more popular business model. Some typical roles in the gig economy include delivery personnel and taxi drivers. In the gig economy, people enjoy a flexible form of work, both from the temporal and spatial dimension. Also, algorithmic management can enhance efficiency and reduce overhead. However, it often lacks labor rights protections, exposing workers to instability, inadequate benefits, and excessive workloads [74]. While flexible, this labor model risks exploiting productivity without offering genuine freedom. Yet, the gig economy is not inherently unsustainable. Much like early industrial workshops evolved into welfare systems, gig economy structures can improve through institutional reforms. Policies such as flexible contracts and tailored social welfare systems for gig workers will be critical. Additionally, advanced technologies can help gig workers develop new skills, address age discrimination, and promote mental health, reducing social inequality.

- Innovation of global collaboration

Simultaneously, AI is reshaping collaboration and innovation paradigms, moving away from stable organizational structures toward more dynamic and flexible cooperation. Future collaborations are likely to involve inter-organizational or individual-to-organization partnerships, driven by AI-enhanced personal capabilities. Empowered individuals can act as independent innovators, fostering network-based partnerships that catalyze innovation across products, industries, and beyond. This trend is evident in fields like academic research, where international collaborations leverage shared resources and diverse ideas for groundbreaking successes. Similar approaches are poised to expand into industries driven by personalized demands, emphasizing efficiency and customization through open innovation environments, adaptive collaboration, and resource optimization.

To summary, we believe that all three changes can bring both challenges and opportunities to our societal development. To maximize their benefit to society, first, these changes need to be carefully studied to learn their positive and negative consequences on the labor market. Next, AI researchers and social scientists should make actionable regulations and plans to minimize the negative impact on humans’ work opportunities. Also, the whole society should work closely to ensure that the social welfare and education systems can evolve to cooperate with the changes in how humans work and collaborate. With the joint efforts from the academia, industry, and government, the appropriate application of AI will be an accelerator to a more productive society and further enhance human well-being.

How will AI transform research methodologies in the social sciences, and what new insights might it enable?

Social scientists strive to uncover predictive patterns in individual and collective behavior. Auguste Comte, often regarded as the "father of sociology," defined social science as "a science of social phenomenon, subject to natural invariable laws, the discovery of which is the object of investigation." [75] This quest for empirically falsifiable laws governing human society has been heavily influenced by the success of "mathematizing" natural sciences, especially physics, since the 19th century. However, this approach has been criticized by interpretivist social scientists [76], who argue that researchers inevitably simplify the nuances of human behavior and social interaction. They also contend that it is challenging for human researchers to remain entirely objective when interacting with human subjects.

In recent decades, computational social science has emerged on top of the increasingly available large-scale behavioral data [77]. This paradigm, fueled by the proliferation of smart devices, has enabled social scientists to reduce their presence during data collection, and improve the granularity and scale of empirical analyses. Modern AI techniques are naturally synergistic because of their data-driven nature [78]. However, the trend of adopting AIs also gives rise to significant barriers, including limited access to high-quality datasets, as well as the increasing costs of training and deploying AI models. These challenges hinder broader engagement of individual researchers.

The breakthrough of LLMs represents a unique opportunity to advance social science research. Pre-trained on web-scale text corpora, LLMs are considered as a "blurry JPEG of the web," [79] offering a compressed representation of vast amounts of online content. Therefore, researchers can potentially improve their data access by querying these large models. Besides, LLMs can potentially be used to automate data analysis process with their emergent ability of deliberate reasoning [80]. Specifically, we have identified three important problems in advanced social science research methodology with modern AI models.

- Synthesizing Research Data

LLMs hold significant promises such as low-cost, scalable data sources, serving as an alternative to traditional methods of collecting data from human subjects. Recent studies have demonstrated the ability of LLMs to simulate the preferences and behavior patterns of certain demographic groups [81]. This capability, often referred to as role-playing, likely stems from LLMs' exposure to diverse online content [82]. By incorporating techniques such as memory mechanisms [83], the quality of synthesized data can be further improved. However, this approach also poses several new challenges. The extent to which LLM-generated responses accurately reflect real human participants is debatable. Studies have highlighted inherent biases in LLMs, such as underrepresenting minority opinions or reducing the variability and nuance of human interaction to oversimplified stereotypes [84]. Researchers find integrating in-depth interviews into LLM's memory can improve the personalization and authenticity of its responses [85], yet achieving a balance between data quality and cost remains a critical challenge.

- Automating Data Analysis

LLMs have demonstrated remarkable zero-shot capabilities in stance detection, sentiment analysis, and free-text response coding, and many other tasks that are central to computational social science [86]. Moreover, techniques like in-context learning allow researchers to customize LLMs for specific tasks using a small number of examples within prompts, eliminating the need for extensive computational resources [87]. These features can reduce barriers to leveraging large-scale data in research. However, their broad applicability also poses ethical challenges, such as unethical and hyper-realistic misinformation generation [88]. Researchers find model alignment can improve LLM's ability to decline unethical queries [89]. Besides, watermarking LLM's output may allow stakeholders to trace the origin of problematic content [90]. However, ensuring responsible use of LLMs should not and cannot be a pure technical problem. It also requires expertise from multiple disciplines and broader societal engagement.

- Exploring Novel Social Theories

A central goal of many social science problems is to develop theories that explain and predict social phenomena. In natural sciences, AI has already proven useful in discovering novel mathematical solutions [91], chemical compounds [92], and protein designs [93]. This success raises an intriguing question of whether LLMs can similarly contribute to social science by generating novel hypotheses or even social theories. While some researchers argue that LLMs can propose innovative hypotheses [94], concerns persist about their originality and validity. Social science theories sometime cannot be easily falsified, requiring nuanced interpretation of human researchers. Thus, the challenge lies in fostering effective human-AI collaboration, where LLMs should serve as complementary tools to aid social scientists in formulating and searching for meaningful theories.

In conclusion, modern AI models, particularly LLMs, offer transformative potential for social science research in terms of synthesizing research data, automating data analysis, exploring novel social theory, and more. However, it is also important to consider the field beyond techniques. It requires delicate balance between AI output and human response, AI automation and human oversight, AI search and human intuition. By fostering human-AI collaboration, this field can unlock new insights while safeguarding rigor and integrity.

Modern AI models, particularly LLMs, offer transformative potential for social science research in terms of synthesizing research data, automating data analysis, exploring novel social theory, and more.

How should regulatory frameworks evolve to govern AI development responsibly and foster global cooperation?

With the rapid advancement of LLMs, generative AI is bringing transformative changes to human society across sectors such as finance, healthcare, and education. At the same time, the unprecedented pace of technological development creates both legal and ethical challenges amidst fierce market competition and uncontrolled growth of the application of this novel technology. These challenges include the dissemination of misinformation and disinformation that threaten democracy and raise public panic, deepfake for malicious conduct, and amplification of societal biases and discrimination. Additionally, as new technologies are increasingly fast iterated and become more and more complex, it brings immense challenges whether they can be regulated within the existing legal frameworks and thus whether a brand new governance framework needs to be established. Striking the right balance between "promoting technological innovation" and "strictly regulating to mitigate societal risks" further exacerbates the difficulties of AI regulation especially when generative AI models are closely related to the power of nations.

From a societal perspective, timely and forward-looking AI policies help ensure that technological advancements align with the baseline ethical values and principles, including protecting fundamental rights, inclusiveness, security and safety, transparency and accountability [95]. For example, the European Union, the United States, and China have introduced laws and regulations seeking to mitigate AI risks. Technology companies such as Microsoft have voluntarily published its AI Principles [96] and established industry forums to recognize the importance of safe and responsible AI development and are committed to making it a reality [97]. Internally, some companies have released responsible AI policies and guidelines that govern their own development and deployment of AI models, thereby demonstrating their social responsibility and helping them in their efforts to garner trust [96] [98]. These efforts aim to mitigate potential harm to individuals and organizations, ensuring that AI serves humanity in a beneficial and controlled manner.

Although governments and technology companies have already taken actions in AI governance, challenges remain. At a macro-level, we noticed that there is a gap between the technology industry and government, where the lack of communication between both parties leads to inconsistent understanding of the benefit and the risk of the existing and future AI technology. Furthermore, the governments of different countries also do not reach consensus on AI regulation. These two challenges lead to different regulations made by different countries, which increase the difficulties of compliance and may possibly increase the cost of AI technology innovation. Beside the macro-level challenges, we also realize challenges at the micro-level. Due to the emergence and complicated nature of LLMs, certain risks might not be identified promptly by developers and regulators during their testing but only be noticed by users in real practice. An example is the identification of LLM jailbreaking prompts, where users can spot that the inappropriate responses to their prompts pass through the content filtering mechanism occasionally [99]. Such nature of LLMs raises the barrier of making comprehensive and timely regulations by both the technology industry and governments.

To address these challenges, we would like to call for actions to ensure that AI technologies advance in a human-centered, safe, secure and trustworthy manner, First, from the top-down perspective, partnerships among AI industry, academia and policymakers should be fostered to address broad AI regulatory challenges. Policymakers, industry practitioners, and researchers in academia must enhance mutual understanding and reinforce the regulatory framework to respect humans' basic values while boosting the advancement of technology. An example is the collaboration between HCI researchers and policymakers to enhance the effective regulation application in research and leverage the results to enhance policymaking [100]. At the same time, the regulations should not focus on the existing but be forward-looking to handle foreseeable challenges. We believe an important practice toward this end is to enhance the collaboration between computer scientists and social scientists. It is possible to combine the in-depth understanding of AI technology development and careful consideration of human society nature through this way. For example, we should learn from the history of human society's development in the past industrial revolutions. Then, with a keen awareness of the differences led by AI technology and other machines, it is able to consider how to regulate AI technology applications to eliminate potentially similar negative effects on our society, such as unemployment, before the damage has been made.

With a keen awareness of the differences led by AI technology and other machines, it is able to consider how to regulate AI technology applications to eliminate potentially similar negative effects on our society, such as unemployment, before the damage has been made.

Furthermore, to ensure that governance can timely address the evolving challenges that AI brings, the governance framework should be designed in both a top-down manner and a bottom-up approach. The evolution of AI governance should be considered as the responsibility of not only the regulator, but every developer and user of technology and the whole community. For example, the frontiers in AI research and development and novice AI users should report their spotted issues whenever in the lifecycle of AI models. These issues should be handled transparently, and corresponding measures need to be added to regulations or guidelines promptly. To boost the bottom-up process, the government and technology companies should construct effective channels to collect these instances and build an in-time mechanism to investigate the risks and update regulations. Furthermore, it is important to leverage social computing techniques to monitor and trace the needs for and discussion about regulations in specific communities, such as mental health [101] and creativity industry [102]. Such dynamics might be considered for timely and targeted policymaking.

Bibliography

[1] D. Kiela et al., "Dynabench: Rethinking Benchmarking in NLP" in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

[2] B. Smith and C. A. Browne, *Tools and Weapons: The Promise and the Peril of the Digital Age*, Penguin, 2021.

[3] I. R. McKenzie, "Inverse Scaling: When Bigger Isn't Better," *Transactions on Machine Learning Research*, 2023.

[4] U. Anwar et al., "Foundational Challenges in Assuring Alignment and Safety of Large Language Models," *Transactions on Machine Learning Research*, 2024.

[5] R. Greenblatt et al., "Alignment Faking in Large Language Models," *arXiv preprint arXiv:2412.14093*, 2024.

[6] J. Kaplan et al., "Scaling Laws for Neural Language Models," *arXiv preprint arXiv:2001.08361*, 2020.

[7] R. Bommasani et al., "On the Opportunities and Risks of Foundation Models," *arXiv preprint arXiv:2108.07258*, 2021.

[8] L. Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback," in *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2022.

[9] I. Gabriel, "Artificial Intelligence, Values, and Alignment," *Minds and Machines*, vol. 30, no. 3, pp. 411-437, 2020.

[10] S. Duan, X. Yi, P. Zhang, T. Lu, X. Xie, and N. Gu, "DENEVIL: Towards Deciphering and Navigating the Ethical Values of Large Language Models via Instruction Learning," in *Proceedings of the International Conference on Learning Representations*, 2024.

[11] J. Yao, X. Yi, X. Wang, J. Wang, and X. Xie, "From Instructions to Intrinsic Human Values: A Survey of Alignment Goals for Big Models," *arXiv preprint arXiv:2308.12014*, 2023.

[12] X. Wang et al., "On the Essence and Prospect: An Investigation of Alignment Approaches for Big Models," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2024.

[13] H. Jiang, X. Yi, Z. Wei, S. Wang, and X. Xie, "Raising the Bar: Investigating the Values of Large Language Models via Generative Evolving Testing," *arXiv preprint arXiv:2406.14230*, 2024.

[14] X. Yi, J. Yao, X. Wang, and X. Xie, "Unpacking the Ethical Value Alignment in Big Models," *arXiv preprint arXiv:2310.17551*, 2023.

[15] J. Yao, X. Yi, X. Wang, Y. Gong, and X. Xie, "Value Fulcra: Mapping Large Language Models to the Multidimensional Spectrum of Basic Human Values," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2024.

[16] J. Yao, X. Yi, and X. Xie, "CLAVE: An Adaptives Framework for Evaluating Values of LLM Generated Responses," in *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2024.

[17] T. Shen et al., "Large Language Model Alignment: A Survey," *arXiv preprint arXiv:2309.15025, 2023*.

[18] W. Wang et al., "Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2024.

[19] C. Li, M. Chen, J. Wang, S. Sitaram, and X. Xie, "CultureLLM: Incorporating Cultural Differences into Large Language Models," in *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2024.

[20] C. Li, D. Teney, L. Yang, Q. Wen, X. Xie, and J. Wang, "CulturePark: Boosting Cross-cultural Understanding in Large Language Models," in *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2024.

[21] R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting, and N. Liu, "Large Language Models in Health Care: Development, Applications, and Challenges," *Health Care Science*, vol. 2, no. 4, pp. 255-263, 2023.

[22] Y. Li, S. Wang, H. Ding, and H. Chen, "Large Language Models in Finance: A Survey," in *Proceedings of the ACM International Conference on AI in Finance*, 2023.

[23] L. Yan et al., "Practical and Ethical Challenges of Large Language Models in Education: A Systematic Scoping Review," *British Journal of Educational Technology*, vol. 55, no. 1, pp. 90-112, 2024.

[24] B. C. Das, M. H. Amini, and Y. Wu, "Security and Privacy Challenges of Large Language Models: A Survey," *ACM Computing Surveys*, 2024.

[25] B. Peng et al., "Securing Large Language Models: Addressing Bias, Misinformation, and Prompt Attacks," *arXiv preprint arXiv:2409.08087*, 2024.

[26] S. Li et al., "Pre-Trained Language Models for Interactive Decision-Making," in *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2022.

[27] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A Survey on Large Language Model (LLM) Security and Privacy: The Good, the Bad, and the Ugly," *High-Confidence Computing*, vol. 4, no. 2, p. 100211, 2024.

[28] Y. Sun, J. He, L. Cui, S. Lei, and C. T. Lu, "Exploring the Deceptive Power of LLM-Generated Fake News: A Study of Real-World Detection Challenges," *arXiv preprint arXiv:2403.18249*, 2024.

[29] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and Transferable Adversarial Attacks on Aligned Language Models," *arXiv preprint arXiv:2307.15043*, 2023.

[30] J. Yi et al., "Defending ChatGPT Against Jailbreak Attack via Self-reminders," *Nature Machine Intelligence*, vol. 5, no. 12, pp. 1486--1496, 2023.

[31] A. Smith, *The Wealth of Nations*, 1776.

[32] "Microsoft Introduces a More Personalized Copilot with Voice and Vision Features," 2024. [Online]. Available: https://news.microsoft.com/source/asia/2024/10/09/microsoft-introuces-a-more-personalized-copilot-with-voice-and-vision-features/

[33] N. Sambasivan, and R. Veeraraghavan, "The Deskilling of Domain Expertise in AI Development," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2022.

[34] D. De Cremer and G. Kasparov, "AI Should Augment Human Intelligence, Not Replace It," *Harvard Business Review*, vol. 18, no. 1, 2021.

[35] A. Sellen and E. Horvitz, "The Rise of the AI Co-Pilot: Lessons for Design from Aviation and Beyond," *Communications of the ACM*, vol. 67, no. 7, pp. 18-23, 2024.

[36] D. Wang et al., "From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2020.

[37] H. Li, Y. Wang, and H. Qu, "Where Are We So Far? Understanding Data Storytelling Tools from the Perspective of Human-AI Collaboration," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024.

[38] B. J. Biddle, *Role Theory: Expectations, Identities, and Behaviors*, Academic Press, 1979.

[39] Y. Chang et al., "A Survey on Evaluation of Large Language Models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1-45, 2024.

[40] K. Zhou et al., "Don't Make Your LLM An Evaluation Benchmark Cheater," *arXiv preprint arXiv:2311.01964*, 2023.

[41] L. Zhou et al., "Predictable Artificial Intelligence," *arXiv preprint arXiv:2310.06167*, 2023.

[42] V. Balachandran et al., "Eureka: Evaluating and Understanding Large Foundation Models," *arXiv preprint arXiv:2409.10566*, 2024.

[43] X. Wang et al., "Evaluating General-Purpose AI with Psychometrics," *arXiv preprint arXiv:2310.16379*, 2023.

[44] K. Zhu, J. Chen, J. Wang, N. Z. Gong, D. Yang, and X. Xie, "DyVal: Dynamic Evaluation of Large Language Models for Reasoning Tasks," in *Proceedings of the International Conference on Learning Representations*, 2024.

[45] K. Zhu, J. Wang, Q. Zhao, R. Xu, and X. Xie, "DyVal 2: Dynamic Evaluation of Large Language Models by Meta Probing Agents," *Proceedings of Machine Learning Research*, 2024.

[46] A. Zou et al., "Representation Engineering: A Top-Down Approach to AI Transparency," *arXiv preprint arXiv:2310.01405*, 2023.

[47] H. Zhao et al., "Explainability for Large Language Models: A Survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 2, pp. 1-38, 2024.

[48] C. Singh, J. P. Inala, M. Galley, R. Caruana, and J. Gao, "Rethinking Interpretability in the Era of Large Language Models," *arXiv preprint arXiv:2402.01761*, 2024.

[49] M. Yuksekgonul, M. Wang, and J. Zou, "Post-Hoc Concept Bottleneck Models," in *Proceedings of the International Conference on Learning Representations*, 2022.

[50] C. Guan, X. Wang, Q. Zhang, R. Chen, D. He, and X. Xie, "Towards A Deep and Unified Understanding of Deep Neural Models in NLP," *Proceedings of Machine Learning Research*, 2019.

[51] B. Kim et al., "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," *Proceedings of Machine Learning Research*, 2018.

[52] M. Li et al., "Evaluating Readability and Faithfulness of Concept-based Explanations," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2024.

[53] C. Chen et al., "Masked Thought: Simply Masking Partial Reasoning Steps Can Improve Mathematical Reasoning Learning of Language Models," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2024.

[54] J. Zhang, X. Wang, Y. Jin, C. Chen, X. Zhang, and K. Liu, "Prototypical Reward Network for Data-Efficient Model Alignment," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2024.

[55] Z. Xu, R. Huang, C. Chen, and X. Wang, "Uncovering Safety Risks of Large Language Models through Concept Activation Vector," in *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2024.

[56] M. Wang et al., "Knowledge Mechanisms in Large Language Models: A Survey and Perspective," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2024.

[57] Z. Chen, K. Zhou, W. X. Zhao, J. Wang, and J. R. Wen, "Not Everything is All You Need: Toward Low-Redundant Optimization for Large Language Model Alignment," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2024.

[58] L. Gao et al., "Scaling and Evaluating Sparse Autoencoders," *arXiv preprint arXiv:2406.04093*, 2024.

[59] T. Lieberum et al., "Gemma Scope: Open Sparse Autoencoders Everywhere All at Once on Gemma 2," in *Proceedings of the Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2024.

[60] X. Wang, K. Wen, Z. Zhang, L. Hou, Z. Liu, and J. Li, "Finding Skill Neurons in Pre-trained Transformer-based Language Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2022.

[61] Darwin, D. Rustin, N. Mukminatien, N. Suryati, E. D. Laksmi and Marzuki, "Critical Thinking in the AI Era: An Exploration of EFL Students' Perceptions, Benefits, and Limitations", *Cogent Education*, vol. 11, no. 1, p. 2290342, 2024.

[62] I. M. Putra and F. I. Abdunnafi, "Understanding the Unexpected: The Adverse Influence of Large Language Models on Critical Thinking and Professional Skepticism in Accounting Education," *Available at SSRN 4914889*, 2024.

[63] Y. Guo and D. Lee, "Leveraging ChatGPT for Enhancing Critical Thinking Skills," *Journal of Chemical Education*, vol. 100, no. 12, pp. 4876-4883, 2023.

[64] T. Adewumi et al., "ProCoT: Stimulating Critical Thinking and Writing of Students through Engagement with Large Language Models (LLMs)," *arXiv preprint arXiv:2312.09801*, 2023.

[65] A. Toner-Rodgers, "Artificial Intelligence, Scientific Discovery, and Product Innovation," *arXiv preprint arXiv:2412.17866*, 2024.

[66] A. Narayanan and S. Kapoor, *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*, Princeton University Press, 2024

[67] J. Butler et al., Eds., "Microsoft New Future of Work Report 2024," Microsoft Research Tech Report MSR-TR-2024-56, 2024. [Online]. Available: https://aka.ms/nfw2024/

[68] M. Weber, *Economy and Society: An Outline of Interpretive Sociology*, University of California Press, 1978.

[69] M. Foucault and A. Kremer-Marietti, *The Archaeology of Knowledge*, Paris: Gallimard, 1995.

[70] K. Mannheim, *Ideology and Utopia: An Introduction to the Sociology of Knowledge*, 2013

[71] S. M. Lipset, *Political Man: The Social Bases of Politics Garden City*, Doubleday, 1960.

[72] R. Florida, *The Rise of the Creative Class*, Brilliance Audio, 2002.

[73] J. A. Robinson and D. Acemoglu, *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*, London: Profile, 2012.

[74] L. Chen, "Labor Order under Digital Control: Research on Labor Control of Take-out Platform Riders," *The Journal of Chinese Sociology*, vol. 9, no. 1, p. 17, 2022.

[75] A. Comte, *The Positive Philosophy of Auguste Comte*, Blanchard, 1858.

[76] P. Rabinow and W. M. Sullivan, Eds., *Interpretive Social Science: A reader*, University of California Press, 1979.

[77] D. Lazer et al., "Computational Social Science," *Science*, vol. 323, no. 5915, pp. 721-723, 2009.

[78] A. Gefen, L. Saint-Raymond, and T. Venturini, "AI for Digital Humanities and Computational Social Sciences," in *Reflections on Artificial Intelligence for Humanity*, 2021, pp. 191-202.

[79] T. Chiang, "ChatGPT Is A Blurry JPEG of the Web," 2023. [Online]. Available: https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web/

[80] J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," in *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2022.

[81] L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate, "Out of One, Many: Using Language Models to Simulate Human Samples," *Political Analysis*, vol. 31, no. 3, pp. 337-351, 2023.

[82] M. Shanahan, K. McDonell, and L. Reynolds, "Role Play with Large Language Models," *Nature*, vol. 623, no. 7987, pp. 493-498, 2023.

[83] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," in *Proceedings of the Annual ACM Symposium on User Interface Software and Technology*, 2023.

[84] M. Cheng, T. Piccardi, and D. Yang, "CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2023.

[85] J. S. Park et al., "Generative Agent Simulations of 1,000 People," *arXiv preprint arXiv:2411.10109*, 2024.

[86] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, and D. Yang, "Can Large Language Models Transform Computational Social Science?," *Computational Linguistics*, vol. 50, no. 1, pp. 237-291, 2024.

[87] M. Hahn and N. Goyal, "A Theory of Emergent In-Context Learning as Implicit Structure Induction," *arXiv preprint arXiv:2303.07971*, 2023.

[88] C. Chen and K. Shu, "Combating Misinformation in the Age of LLMs: Opportunities and Challenges," *AI Magazine*, vol. 45, no. 3, pp. 354-368, 2024.

[89] Y. Liu et al., "Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment," in *Proceedings of the International Conference on Learning Representations*, 2024.

[90] S. Dathathri et al., "Scalable Watermarking for Identifying Large Language Model Outputs," *Nature*, vol. 634, no. 8035, pp. 818-823, 2024.

[91] B. Romera-Paredes et al., "Mathematical Discoveries from Program Search with Large Language Models," *Nature*, vol. 625, no. 7995, pp. 468-475, 2024.

[92] D. A. Boiko, R. MacKnight, B. Kline, and G. Gomes, "Autonomous Chemical Research with Large Language Models," *Nature*, vol. 624, no. 7992, pp. 570-578, 2023.

[93] Z. Zheng, Y. Deng, D. Xue, Y. Zhou, F. Ye, and Q. Gu, "Structure-informed Language Models Are Protein Designers," *Proceedings of Machine Learning Research*, 2023.

[94] R. Wang, E. Zelikman, G. Poesia, Y. Pu, N. Haber, and N. D. Goodman, "Hypothesis Search: Inductive Reasoning with Language Models," in *Proceedings of the International Conference on Learning Representations*, 2023.

[95] "OECD AI Principles Overview," [Online]. Available: https://oecd.ai/en/ai-principles/

[96] "The Microsoft Responsible AI Standard," [Online]. Available: https://www.microsoft.com/en-us/ai/principles-and-approach/?msockid=1f56b30ebe61621830fca7a2bf98637b/

[97] "Frontier Model Forum: Advancing Frontier AI Safety," [Online]. Available: https://www.frontiermodelforum.org/

[98] "Advancing Bold and Responsible Approaches to AI," [Online]. Available: https://publicpolicy.google/responsible-ai/

[99] X. Shen, Z. Chen, M. Backes, Y. Shen, and Y. Zhang, "'Do Anything Now': Characterizing and Evaluating In-the-Wild Jailbreak Prompts on Large Language Models," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2024.

[100] Q. Yang et al., "The Future of HCI-Policy Collaboration," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024.

[101] E. Sharma and M. De Choudhury, "Mental Health Support and its Relationship to Linguistic Accommodation in Online Communities," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2018.

[102] Q. Guo, K. Yuan, C. He, Z. Peng, and X. Ma, "Exploring the Evolvement of Artwork Descriptions in Online Creative Community under the Surge of Generative AI: A Case Study of DeviantArt," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 2024.

Acknowledgments

This whitepaper is a joint effort by the Societal AI team[8] in Microsoft Research Asia (MSRA) and a wide range of computer and social scientists from the industry and academia, Tianguang Meng and Fengli Xu from Tsinghua University, Jindong Wang from College of William and Mary, Xiting Wang from Renmin University of China, Linus Huang, an MSRA StarTrack Scholar from the Hong Kong University of Science and Technology, Lexin Zhou, an MSRA StarBridge Scholar, and Vivian Ding from Microsoft CELA. Other contributors to the whitepaper include Jianjun Yu from Tsinghua University, and Binghao Huan from Peking University, Hao Yang from Nanjing University, and Bryan Xu from New York University, who contributed to the whitepaper during their internships in MSRA.

The publication of this whitepaper is also inseparable from the outstanding designer team led by Yang Ou and Scarlett Li from MSRA. The team also has Bella Guo and two intern designers: Tianmeng Liu from University of Washington and Yiqi Tan from Capital Normal University.

In addition to the contributors, we would like to express our deepest gratitude to all researchers who have supported the Societal AI initiative through their long-term and fruitful collaboration with the MSRA team[9], insightful talks and lively discussions in workshops and seminars[10], and inspiring and in-depth lectures on Societal AI topics[11].

Finally, we would like to acknowledge Lidong Zhou from MSRA for his continuous warmest support and invaluable suggestions to not only the whitepaper but also the whole Societal AI team and project.

---

[08] https://www.microsoft.com/en-us/research/project/societal-ai/

[09] https://www.microsoft.com/en-us/research/project/societal-ai/collaboration/

[10] https://www.microsoft.com/en-us/research/project/societal-ai/events/

[11] https://www.microsoft.com/en-us/research/project/societal-ai/lectures/

© 2025 Microsoft Corporation. All rights reserved.