Information Processing and Management 61 (2024) 103665

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

AI for social science and social science of AI: A survey

Ruoxi Xu [a,b], Yingfei Sun [a], Mengjie Ren [b], Shiguang Guo [b], Ruotong Pan [b], Hongyu Lin [b,*], Le Sun [b,c], Xianpei Han [b,c]

[a] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing, China
[b] Chinese Information Processing Laboratory, Institute of Software, Chinese Academy of Sciences, Beijing, China
[c] State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

Keywords:
Social science
Large language models
AI simulation

ABSTRACT

Recent advancements in artificial intelligence, particularly with the emergence of large language models (LLMs), have sparked a rethinking of artificial general intelligence possibilities. The increasing human-like capabilities of AI are also attracting attention in social science research, leading to various studies exploring the combination of these two fields. In this survey, we systematically categorize previous explorations in the combination of AI and social science into two directions that share common technical approaches but differ in their research objectives. The first direction is focused on *AI for social science*, where AI is utilized as a powerful tool to enhance various stages of social science research. While the second direction is the *social science of AI*, which examines AI agents as social entities with their human-like cognitive and linguistic capabilities. By conducting a thorough review, particularly on the substantial progress facilitated by recent advancements in large language models, this paper introduces a fresh perspective to reassess the relationship between AI and social science, provides a cohesive framework that allows researchers to understand the distinctions and connections between AI for social science and social science of AI, and also summarizes state-of-art experiment simulation platforms to facilitate research in these two directions. We believe that with the ongoing advancement of AI technology and the increasing integration of intelligent agents into our daily lives, the significance of the combination of AI and social science will become even more prominent.

1. Introduction

Building machines that can think, learn and create is the fundamental pursuit of artificial intelligence (AI) (Russell, 2010). How to develop machines with general intelligence comparable to, or even greater than, that of human beings has never lost its appeal (Goertzel, 2014). Recently, significant advancements have been made in the AI field (Zhao et al., 2023), particularly with the emergence of large language models (LLMs) such as ChatGPT and GPT-4 (OpenAI, 2023b). These developments have led to the rethinking of the possibilities of artificial general intelligence (AGI) (Zhao et al., 2023).

The increasing human-like capabilities of AI are also attracting attention in social science research. There have been numerous studies exploring the combination of AI and social science (Bail, 2023; Chen, 2023a; Ziems et al., 2023). Along these lines, many novel research directions have been explored, including research tasks proposing (Banker, Chatterjee, Mishra, & Mishra, 2023; Park, Kaplan et al., 2023), social science simulation (Brand, Israeli, & Ngwe, 2023; Chu, Andreas, Ansolabehere, & Roy, 2023; Kjell, Kjell, & Schwartz, 2023), AI agent governing (Jiang, Zhang, Cao, Kabbara, & Roy, 2023; Li et al., 2023; Miotto, Rossberg, & Kleinberg, 2022) and so on.

* Corresponding author.
E-mail addresses: ruoxi2021@iscas.ac.cn (R. Xu), hongyu@iscas.ac.cn (H. Lin).

https://doi.org/10.1016/j.ipm.2024.103665
Received 28 June 2023; Received in revised form 27 October 2023; Accepted 16 January 2024
Available online 8 February 2024

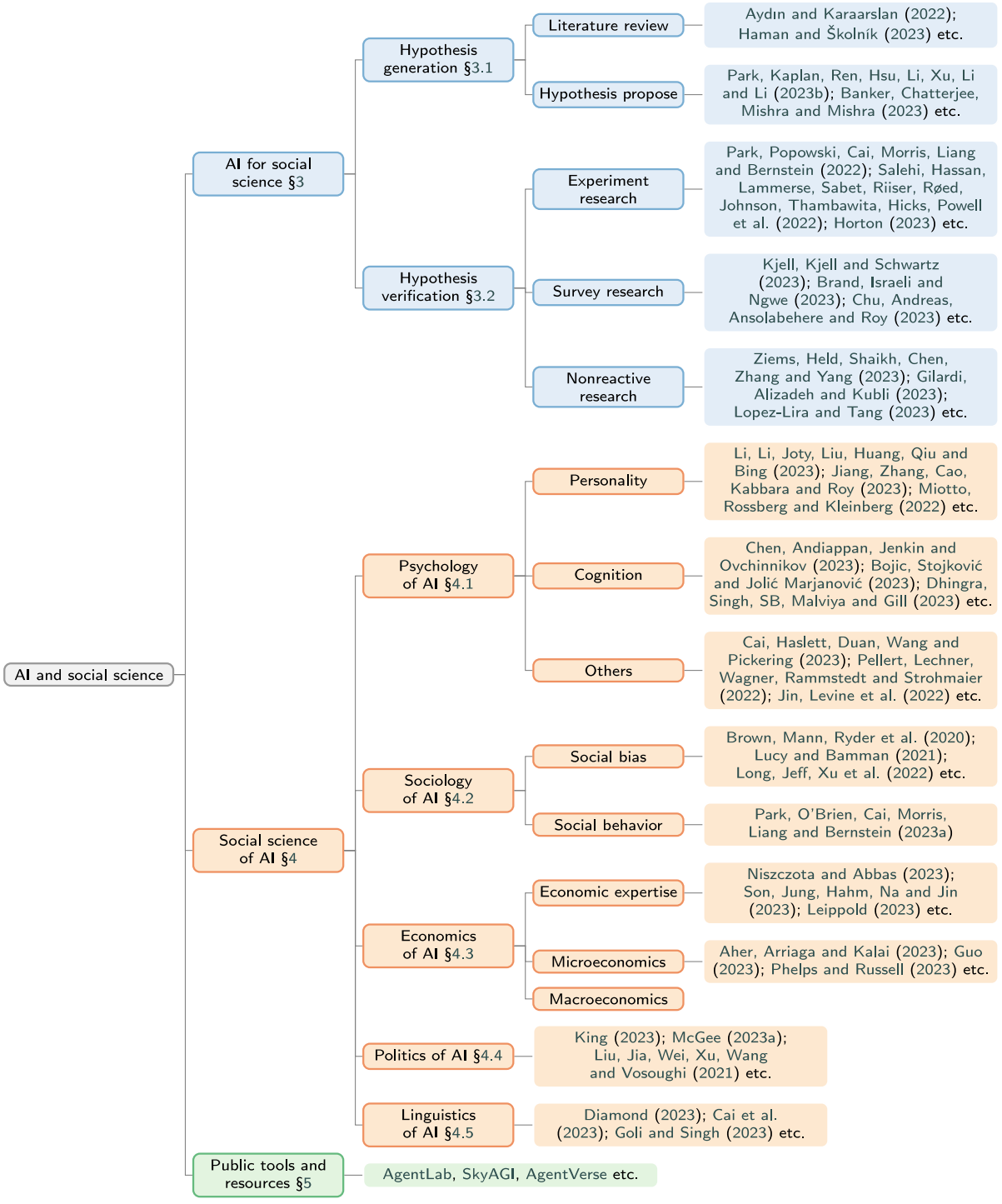

Fig. 1. Overview of the intersection of AI and social science. We have separately discussed "AI for social science" which summarizes the application of AI at every stage of social science research to provide guidance on tool selection for researchers, "social science of AI" which systematically describes the intelligence level and characteristics of AI agents from a social science perspective on different sub-disciplines, and "Public tools and resources" which focuses on simulation tools. These fields share technical methodologies to some extent, yet they possess distinct research subjects and objectives.

Despite the large number of related studies, the existing studies tend to concentrate on specific instances of AI and social science intersection, thereby lacking a unified perspective to effectively distinguish and outline AI's role in social science research and its own social characteristics. In reality, the combination of AI and social science can be divided into two distinct directions. On the

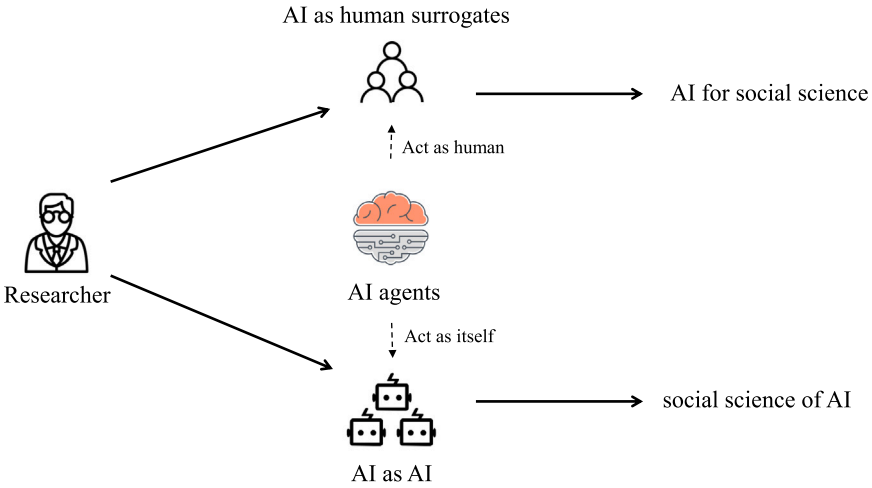

Fig. 2. Computer simulation respectively in the context of "AI for social science" and "social science of AI". For "AI for social science", AI agents are deployed to mimic human behaviors to enhance the understanding of human society. Conversely, "social science of AI" delves into AI agents' own social questions.

one hand, the superior performance of AI allows them to serve as effective tools for social science research, such as using AI for literature searching and reviewing (McGee, 2023b), proposing questions and hypotheses (Banker et al., 2023; Park, Kaplan et al., 2023), analyzing data (Ziems et al., 2023), assisting with writing (Chen, 2023b; Dergaa, Chamari, Zmijewski, & Saad, 2023), and more. Systematically outlining the potential applications of AI in different stages of social science research can provide a valuable guide for researchers to choose appropriate research tools. We refer to this direction as *AI for social science* in this paper. On the other hand, just as early myths and parables emphasized the social and ethical questions around human-created intelligence (Kieval, 1997; McCorduck & Cfe, 2004; Pollin, 1965), today's intelligent machines present their own interesting social questions (Frank, Wang, Cebrian, & Rahwan, 2019) and expanding research starts to explore and understand AI agents as social entities. Particularly, current AI agents, especially large language model-based agents, are exhibiting cognitive, logical reasoning, and linguistic capabilities on par with or even surpassing those of humans, along with unique behavioral characteristics (OpenAI, 2023b). Communities constituted by AI agents also exhibit emergent behaviors similar to human societies (Park, O'Brien et al., 2023). This provides an interesting case for attempting to extend the social science to more universal phenomena of machines (Klein & Kleinman, 2002) and also presents a valuable opportunity to reevaluate a fundamental axiom in social science: human behavior can be understood as possessing unique social characteristics (Woolgar, 1985). Exploring AI from a social science perspective can also provide crucial insights and guidance to make AI development more congruent with societal needs and human values. We refer to this direction as *social science of AI* in this paper. There is an important point to note, the term "social science" as used in this paper extends its traditional definition. It is used in a broader sense to provide a research perspective for describing certain high-level behaviors of humans or models, rather than equating it with actual human social behaviors.

Although these directions share common technical approaches, they have distinct research objectives, significance, and scopes of application. For example in Fig. 2, using AI agents for simulation serves as a technical method that could be applied to both directions, but with different objectives. When used for the former, researchers' aim is to align the behavior of AI agents with human behavior as closely as possible, in order to study the operational laws of human society in a cost-effective, fast, and ethically risk-averse manner (Park et al., 2022; Salehi et al., 2022). When used for the latter, the objective is to explore the behavioral laws of AI itself, with a particular focus on its unique aspects, especially those differing from the operational laws of human society (Guo, 2023). The absence of surveys from the above two perspectives makes it hard to ground each work's research significance and application scope, hindering us from comprehension and harnessing of the distinctions and connections between these two directions. A joint analysis of their research progress can help to grasp the big picture of the current state of the combination of AI and social science.

To this end, we conduct a comprehensive review from these two directions respectively. We conduct a joint analysis of their research progress, comparing their similarities and differences to present an overview of the current state of the combination of AI and social science. Considering that recent remarkable advancements in this field can be largely attributed to the development of large language models (Zhao et al., 2023), this paper narrows down its scope to the combination of large language models and social science, approaching the topic from both the angles of AI for social science and social science of AI. The main organization of this survey is summarized in Fig. 1. Specifically, from the angle of AI for social science, we discuss large language models' potential as a highly efficient tool that can be integrated into existing research methodologies, significantly enhancing the efficiency of social science research. To achieve this, we structure the content according to the roles that AI plays in both the stages of hypothesis generation and verification within the social science research process (Bryman, 2016; Donovan & Hoover, 2013). For hypothesis generation, we mainly focus on how AI can help human beings in literature reviewing and hypothesis proposing. For hypothesis verification stage, we respectively examine how large language models function in various research methods such as experiment research, survey research and non-reactive research. From the perspective of social science of AI, we are referring to a broad field of

social science that focuses on regarding large language models as its research subject. We categorize the behavioral studies of these models according to the subfields within social science, following academic categorization. More specifically, we have compiled the behavioral laws of large language models by examining them from the viewpoints of psychology, sociology, economics, politics, and linguistics. Additionally, we have also compiled a summary of currently available tools in this field to facilitate research in the aforementioned areas. These platforms utilize large language models as agents and allow for the setting and implementation of intervention conditions to simulate diverse social situations, interactions, and behaviors. They serve the purpose of simulating human behavior for studying human societies, as well as exploring AI societies, thus catering to both of the above directions.

Generally, we summarize our contributions as follows:

- We present a perspective of revisiting AI and social science combinations from two directions: *AI for social science* and *social science of AI*. We elaborate on the connections and distinctions between these two directions, grounding the research value and application scope of relevant work.
- In light of the significant strides enabled by recent advancements in large language models in these two directions, we conduct a literature review, which summarizes the research landscape, discusses the limitations of the existing research, and sheds light on potential future directions in the combination of AI and social science.
- We collect and compare existing open source large language model-based simulation tools. These platforms can effectively serve as a foundation to facilitate future research in the two directions mentioned above.

The structure of this survey is organized as follows: Section 3 introduces the application of large language models in social science research, while Section 4 delves into social science research that takes large language models as the subject of study. Section 5 provides information about the resources and tools available. Finally, we conclude the survey in Section 6, summarizing the main findings and discussing the remaining issues for future work.

2. Background

Nowadays, the AI community, and even the whole society, is witnessing the significant impacts brought about by large language models. Large language models typically refer to transformer language models that contain hundreds of billions (or more) of parameters (Shanahan, 2022), which are trained on massive text data. Notable examples include GPT-3 (Brown, Mann, Ryder, et al., 2020), PaLM (Chowdhery et al., 2022), LLaMA (Touvron et al., 2023) and OpenAI's ChatGPT, which amassed 100 million users in less than two months, setting a new record in history. These models have exhibited strong capacities to understand natural language and solve complex tasks (via text generation), capturing the attention and imagination of investors, consumers, and organizations.

Improvements in large language models have been so fast, and the potential societal repercussions have been so profound, that a broad cross-disciplinary lens from computer science and social science is necessary to start making sense of the implications. Firstly, the ability of large language models to generate texts for a broad range of tasks via an intuitive natural language interface may hold great promise for social science research. This has opened up avenues for various applications, including literature searching and reviewing (McGee, 2023b), proposing questions and hypotheses (Banker et al., 2023; Park, Kaplan et al., 2023), analyzing data (Ziems et al., 2023), assisting with writing (Chen, 2023b; Dergaa et al., 2023), and more. Secondly, the evolution of large language models has significantly enhanced their capacity to exhibit human-like characteristics, leading to a surge in research regarding language models as representations of human entities (Andreas, 2022; Krishna, Lee, Fei-Fei, & Bernstein, 2022; Park et al., 2022). Research includes exploring the collaborative capabilities of large language models in complex tasks (Irving, Christiano, & Amodei, 2018), developing "generative agents" to investigate emergent social behaviors (Park, O'Brien et al., 2023), employing GPT-3-based agents as substitutes for human participants (Aher, Arriaga, & Kalai, 2023) and so on. Thirdly, the advancements in large language models have prompted a reconsideration of the ethical and societal issues it may entail. These previous works have highlighted the transformative effects brought about by the emergence of large language models in the intersection of AI and social science. Our survey aims to provide a comprehensive overview of these developments and address the existing limitations and potential of this field.

3. AI for social science

AI for social science refers to the application of AI in traditional social science research. Unlike social science of AI, this section emphasizes large language models' human-like intelligence, which can mimic human behavior to help social science research. In this section, we will draw upon the research paradigm outlined in books of Bryman (2016) and Donovan and Hoover (2013), and discuss the application of large language models as multi-purpose tools at every stage of social science research as shown in Fig. 3. This aims to provide a comprehensive and informed perspective for social science researchers on how to apply large language models in the process of their research to enhance efficiency, while also revealing the untapped potential of large language models, warning about potential risks and ethical issues, and indicating possible future directions in their application.

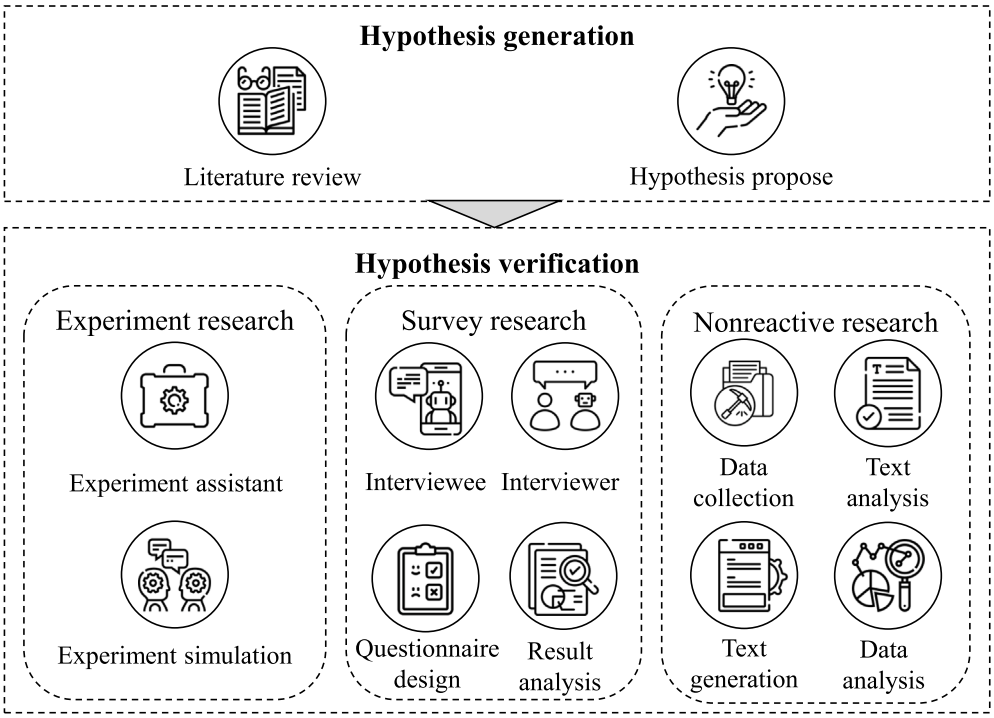

Fig. 3. The application of large language models at every stage of social science research. Large language models offer new possibilities for improving existing social science research processes and automated science, but also bring new potential risks and ethical issues. Social science researchers should carefully consider whether and how to apply large language models in their research.

3.1. Hypothesis generation

Hypothesis generation, which serves as the foundation and initial step of social science research, is the task of mining meaningful implicit associations between unrelated social science concepts (Jha, Xun, Wang, & Zhang, 2019). In the early days, hypothesis generation was largely driven by brainstorming of researchers, drawing inspiration from existing theories, patterns of anomalies in data or cross-disciplinary connections that involved serendipitous discoveries (Jaccard & Jacoby, 2019). However, as the volume of research literature grows, researchers have started to explore quicker and more efficient methods of generating solid hypotheses (Evans & Rzhetsky, 2010; Krenn & Zeilinger, 2020; Wilson et al., 2018). In the following, we will present some attempts to accomplish hypothesis generation tasks using large language models.

Literature review. is the understanding, summarization, and critical thinking about the academic literature on a specific topic. Researchers have been exploring the use of large language models to assist in literature review. Some studies leverage large language models to aid in searching for relevant literature. Wang, Scells, Koopman, and Zuccon (2023) used ChatGPT to formulate and refine boolean queries for systematic reviews, finding that ChatGPT compared favorably with the current state-of-the-art automated boolean query generation methods in terms of precision, at the expenses of a lower recall. The paper also found that guided prompts led to higher effectiveness than single prompt strategies. Other works focus on enabling large language models to read articles and automatically summarize the key points within them. Aydın and Karaarslan (2022) used ChatGPT to paraphrase the abstracts of relevant papers and answer questions to automatically generate literature review, which has been cited over 100 times. Elicit[1] is also one typical application, an AI Research Assistant based on large language models, which can find relevant papers without perfect keyword match, summarize takeaways from the paper specific to your question, and extract key information from the papers. However, it should be noted that directly relying on large language models for literature recommendation and summarization is currently infeasible due to the risk of unreliable papers and related information being generated (Haman & Školník, 2023).

Hypothesis propose. is to propose possible explanations to some phenomenon or event. Researchers from various fields have made attempts to use large language models to help with this task. For example, Park, Kaplan et al. (2023) used GPT-4 to generate scientific hypotheses and drew the conclusion that current large language models seemed to be able to effectively structure vast amounts of scientific knowledge and provide interesting and testable hypotheses while the error rate was high. Banker et al. (2023)

[1] https://elicit.org/

Table 1
The comparisons between traditional methods and large language models as tools at every stage of social science research. The advantages are marked in bold. From the table, we can easily find that although large language models are more advantageous on cost, speed, generality and accessibility across various research stages, the critical current limitations of validity, possible ethical risks and lack of domain knowledge still hinder its real-word applications. Given the rapid advancement of technology, the temporal efficacy of the relative advantages delineated in the table is noteworthy.

Research stages	Traditional methods	Large language models
Hypothesis generation		
Speed	Low	**High**
Validity	**High**	Low
Novelty	Low	**High**
Hypothesis verification		
Experiment research		
Cost	High	**Low**
Speed	Low	**High**
Reproducibility	Low	**High**
Scalability	Low	**High**
Fidelity	**Entire**	Not Sure
Survey research		
Cost	High	**Low**
Engagement	Low	**Entire**
Interaction	Fixed	**Natural**
Bias	**Low**	Not Sure
Nonreactive research		
Generality	Single-purpose	**Multiple-purpose**
Accessibility	Low	**High**
Numerical analysis	**Accurate**	Not Sure

utilized a fine-tuned version of GPT-3 to generate psychological hypotheses and engaged 50 psychology experts to evaluate their quality, revealing that the model's generated hypotheses were not mere replicas of previously generated human hypotheses, and exhibited no significant differences in terms of clarity, impact, and originality compared to human-generated hypotheses. Tang et al. (2023) employed large language models to generate "less likely" hypotheses, effectively assisting humans in comprehensively examining problems and eliminating cognitive biases caused by their own knowledge and experience. At present, the application of large language models in hypothesis generation is still somewhat rudimentary. Strategies for enhancing hypothesis quality include fine-tuning large language models within specific domains (Banker et al., 2023), stepwise questioning (Wei et al., 2022), adversarial dialogue (Park, Kaplan et al., 2023) and so on.

Conclusion. Current research on using large language models for hypothesis generation focuses on exploring its feasibility and validity, and has commonly unveiled promising prospects and potential of employing large language models for this task. The application methods mainly involve interactive design of prompts. Users input the topic they want to review. Based on these inputs, large language models automate the process of formulating and refining boolean queries, extracting core points from the search results, and generating hypotheses about potential relationships among the objects of interest.

Compared to traditional methods, the **advantages** mainly lie in their exceptional performance in language understanding and generation, enabling them to quickly analyze existing research, identify knowledge gaps, and establish connections between seemingly unrelated ideas (Dahmen et al., 2023). This provides them with a natural advantage in language-based disciplines such as psychology (Banker et al., 2023). However, researchers need to be noted that currently, large language models can only be used as auxiliary and inspirational tools in the early stage of research, and have the following **limitations**: (1) Fabricated or incorrect information, which may mislead users. This is because large language models lack of understanding regarding the validity of output content and simply spill out them without clear rationalization (Park, Kaplan et al., 2023). (2) High sensitivity to the prompt, which results in a significant investment of effort in prompt design but yields uncertain outcomes (Wang, Scells et al., 2023). (3) Limited context length, which makes it hard to handle long and multiple documents. In summary, the comparisons between traditional methods and large language models as tools at every stage of social science research are presented in Table 1.

In order to better harness the potential of large language models in hypothesis generation, **future directions** that we may consider include: (1) Integrating specialized domain knowledge, by retrieval augmented techniques or domain-specific training. This would help reduce hallucination. (2) Developing high-reward prompt strategies. This could involve considering novel prompt generation techniques or reward mechanisms to guide the model's hypothesis generation process. (3) Expanding the context windows of the large language models. By allowing the models to consider a larger context, they would have access to more comprehensive information, potentially leading to more robust and insightful hypotheses.

3.2. Hypothesis verification

Once the research topic and hypotheses are established, social science researchers engage in hypothesis verification. This process involves collecting and analyzing data to provide evidence that either supports or refutes the proposed hypotheses (Donovan

& Hoover, 2013). In traditional social science research, hypothesis verification typically falls into quantitative methods like experimental research, survey research and nonreactive research, as well as qualitative methods such as field research and historical-comparative research (Bryman, 2016; Juren Lin, 2017; Yuan, 2013). Given that large language models are currently limited in their applicability to qualitative research, we primarily discuss the role of large language models in quantitative methods.

3.2.1. Experiment research

A laboratory experiment is "an inquiry for which the investigator plans, builds, or otherwise controls the conditions under which phenomena are observed and measured" (Willer & Walker, 2007). The common practice involves manipulating conditions for some research participants while leaving them unaltered for others, aiming to compare the responses across groups to uncover consistent behavioral patterns. In the following, we will explore the applications of large language models in experimental research.

Experiment assistant. refers to the use of large language models in social science experiments to automate some simple but labor-intensive tasks that would normally be done by researchers. For instance, they can assist in creating hypothetical scenarios iteratively with feedback from researchers (Bail, 2023), which can enhance the external validity and comparability of the experimental conditions. Besides, large language models are capable of synthesizing the necessary information for experiments, eliminating the need for the use of real-life information utilization. This safeguards the privacy of individuals whose information could potentially be used in these studies.

Experiment simulation. aims to design a platform to explore, optimize, and predict behaviors of complex systems that might be challenging to investigate in the real world. In simulation experiments, large language models are usually used as believable proxies of human behavior (Aher et al., 2023; Park, O'Brien et al., 2023). For example, Park et al. (2022) provided a typical application where large language models were utilized to simulate the behavior of community users, assisting designers in gaining insights into the various possibilities of social interactions and identifying potential edge cases that could lead to the breakdown of a community. Horton (2023) considered GPT-3 AIs as homo silicus agents, and demonstrated their ability to qualitatively recover findings from three classic behavioral economics experiments with real humans. Guo (2023) designed well-crafted prompts to enable GPT agents to participate in strategic game experiments and achieve realistic performance. Park, O'Brien et al. (2023) constructed a fully large language model-driven simulated community, where they observed human-like individual behaviors and emergent behaviors.

Conclusion. In experimental research, large language models can serve dual roles — they can act as experiment assistants and as believable proxies of human behavior, becoming subjects of the experiment themselves. The latter, in particular, has attracted increasing attention in both AI and social science as large language models are increasingly capable of simulating human-like responses and behaviors. Currently, the design of AI agents is still relatively crude, usually including four modules: profile, memory, planning, and action (Wang, Ma et al., 2023), and warrants further improvement.

Using large language models for simulating experiments offers several **advantages**: (1) Improved efficiency, reduced costs, and enhanced scalability (Bybee, 2023; Guo, 2023). (2) Circumventing the ethical issues associated with human subjects. This opens the door to experiments that may be deemed unethical if performed on humans, such as the classic Stanford Prison Experiment (Zimbardo, Haney, Banks, & Jaffe, 1971).

However, social scientists must also proceed with caution in this area, taking into account the following **limitations**: (1) Uncertain believability. There is now no "gold standard" study demonstrating that groups of automated agents can accurately simulate humans (Bail, 2023). (2) Low transparency and reproducibility. Since large language models themselves are still black boxes, we cannot provide a thorough explanation of their behaviors.

In order to address the aforementioned limitations, potential **future directions** include: (1) developing methods for evaluating the quality of large language model-based simulations, and (2) incorporating insights from cognitive science to guide the development of AI agent frameworks and enhance their behaviors' human-likeness and rationality.

3.2.2. Survey research

Survey research is a fundamental methodology in social science, which uses written questionnaires or formal interviews to collect information on the beliefs, opinions, characteristics, and past or present behaviors of a target group (Bryman, 2016). The core of modern survey research is three key components: sampling, measurement, and analysis (Wright, Marsden, et al., 2010). The following will explore the role that large language models play in each stage of survey research.

Sampling. involves selecting representative samples from human populations, whose observed characteristics provide unbiased estimates of the characteristics of those populations. Large language models present a novel option for sampling, serving as proxies for specific human subgroups. This enables the avoidance of the sampling step, or rather, utilizes the extensive training database of large language models directly as the sample for the study. Existing studies have proposed and explored the possibility of using language models as effective stand-ins for particular human demographics in the realm of social science research. For example, Argyle et al. (2023) compared real human participants from multiple large surveys in the United States and "silicon samples" which were created by conditioning large language models on socio-demographic backstories from them, demonstrating that the "algorithmic bias" within GPT-3 was both fine-grained and demographically correlated. Bisbee, Clinton, Dorff, Kenkel, and Larson (2023) investigated the quality, reliability, and reproducibility of synthetic survey data generated by the popular closed-source large language model, ChatGPT. The experimental results suggested that the average scores generated by ChatGPT closely aligned with the averages from baseline survey (conducted from 2016 to 2020 on the U.S. national elections). However, when it came to more advanced features, such as variance, subgroups, and statistical inferences, it often led researchers to draw conclusions that

differed from those relying on human respondents. Dillion, Tandon, Gu, and Gray (2023) took moral judgments as an example to investigate whether and when language models could potentially replace human participants in psychology. The analysis indicated that language models aligned most closely with humans when the contextual circumstances involved explicit features that drove human judgments, didn't pertain to competitive situations, and when the subjects being judged were representative within the training data. Rao, Leung, and Miao (2023) conducted the Myers-Briggs Type Indicator (MBTI) test on large language models agents of different subpopulations and showcased the ability of ChatGPT in analyzing personalities of different groups of people. Brand et al. (2023) interviewed GPT-3 to estimate customers' willingness-to-pay for products and features and found that large language models could generate responses that aligned with economic theories and consumer behavior patterns. Chu et al. (2023) adapted language models to subpopulation-specific media diets and successfully simulated how the subpopulation would respond to survey questions.

Measurement. focuses on designing questions to draw out valid and reliable responses across a broad spectrum of subjects, which are often characterized as "the art of asking questions". While it is natural to leverage large language models to assist in the design of questionnaires or interview questions, more researchers are focusing on the role large language models play in facilitating a paradigm shift in the measurement methods within survey research — from closed-ended rating scales, to open-ended response questionnaires, and then further towards more natural interactive interviews. For example, Kjell, Sikström, Kjell, and Schwartz (2022) compared the results of psychological surveys using rating scales and natural language-based open-ended questionnaires. The latter were found to achieve accuracy that either exceeded or rivaled the typical methods of reliability in rating scales, which was often considered as the theoretical upper limit. Kjell et al. (2023) provided a future outlook of finer granularity and automated interactive interviews, making full use of interviewees' own words to best elicit truthful responses.

Analysis. is a step using multivariate data analysis techniques to identify and understand the statistical relationships among various variables. Large language models can be used to analyze qualitative data, such as interview responses, to identify patterns, relationships, and common themes (Abbas, 2023). For instance, Yang, Ji, Zhang, Xie and Ananiadou (2023) utilized large language models, specifically ChatGPT, to perform mental health analysis and highlighted the significant potential of large language models in improving the interpretability of mental health analysis. However, large language models, which are not specifically designed for analyzing quantitative data, are currently not the primary method for survey research analysis in social science. Instead, survey data is typically presented in the form of charts, graphs, or tables, and analyzed using statistical methods (Bryman, 2016). Future versions of language models may be able to integrate tools like Python and R libraries to conduct quantitative data analyses (Mialon et al., 2023).

Conclusion. The current applications of large language models in survey research primarily revolve around three main directions: (1) Effective proxies for specific human sub-populations. (2) Interactive interviewers. (3) Result analysis tools.

For the first direction, the current work has demonstrated that proper conditioning will cause large language models to accurately emulate response distributions from a wide variety of human subgroups. This approach can effectively address limitations regarding the number of questions, frequency and the subpopulation due to cost constraints (Chu et al., 2023), as well as the common challenge of low response rates (Bhattacherjee, 2012) in survey research, offering significant advantages in terms of engagement and cost. However, whether and which large language models can truly represent humans remains an open question (Argyle et al., 2023). This approach fundamentally relies on "algorithmic bias", which is heavily influenced by the training data and is susceptible to producing unfair and non-objective results. In light of these considerations, we do not propose that large language models should completely replace traditional sampling methods in survey research. Instead, we see their potential in simulating population responses to assist in survey design. This hybrid approach allows us to harness the strengths of large language models while still recognizing the importance of traditional sampling techniques to maintain the integrity and fairness of survey results.

Several researchers have envisioned the impact of language models on the form of survey. The advantage of large language models lies in their ability to fully utilize the individuals' own language to describe the information needed by researchers, which has the potential not only to gradually improve current assessments but also to fundamentally alter the nature of measuring and describing personal states, ultimately enhancing our understanding of social science. However, utilizing large language models to facilitate measurement also poses risks. Inherent biases in large language models and the potential for data leakage must be carefully navigated when implementing large language models in research scenarios.

For result analysis, large language models are primarily used for text analysis and are seldom employed for numerical analysis because they are not proficient in it. Future research can consider the integration of large language models with specialized computational tools.

3.2.3. Nonreactive research

Nonreactive research refers to the research method where the participants are not aware that their information is part of the study, unlike experiment research and survey research that actively engage the people we study by creating experimental conditions or directly asking questions (Juren Lin, 2017). This method may reduce bias due to interference from researchers or measurement instruments (Trochim & Donnelly, 2001). In this section, we will adhere to the taxonomy in Juren Lin (2017), and explore the roles that large language models play in content analysis and existing statistics analysis.

Content analysis. is a widely used technique for examining the content contained in written documents or other communication media. The remarkable performance of large language models in various traditional NLP tasks has attracted extensive attention about using them in content analysis tasks within the field of social science.

Some social science researchers employ large language models to perform text classification, a basic and important task that involves labeling or categorizing texts according to predefined categories (Aggarwal & Zhai, 2012). Common text classification tasks in the field of social science include: (1) **Sentiment analysis**, which aims to identify and extract the emotional attitudes in the text, such as joy, anger and sorrow. It is a widely applied technique in psychology and political science. In psychology, sentiment analysis helps researchers understand people's emotional states, stress levels, and mental health conditions. Rodríguez-Ibánez, Casánez-Ventura, Castejón-Mateos, and Cuenca-Jiménez (2023) suggested that large language models were the future paradigm for sentiment analysis, due to their zero-shot setting and simple invocation. However, they also pointed out the limitations of GPT-3 in the current tasks. ChatGPT performed excellently in three text-based mental health classification tasks, including stress detection, depression detection, and suicide detection (Lamichhane, 2023). ChatGPT also applied to differentiate paranoid texts from non-paranoid ones in some studies (Uludag, 2023). Rathje et al. (2023) evaluated the performance of GPT-3.5 and GPT-4 on multilingual sentiment and discrete emotions tasks and found that in many cases, GPT models performed close to (sometimes better than) fine-tuned machine learning models. They argued that GPT models offered a promising avenue for cross-lingual research in psychology. Wu, Irsoy et al. (2023) introduced BloombergGPT, a 50 billion parameter language model tailored for the financial domain, and evaluated it on two financial sentiment analysis datasets FPB (Malo, Sinha, Korhonen, Wallenius, & Takala, 2013) and FiQA SA (Maia et al., 2018). BloombergGPT outperformed general models such as GPT-Neo (Black, Leo, Wang, Leahy, & Biderman, 2021), OPT (Zhang et al., 2022) and BLOOM (Scao et al., 2023) on both tasks. Frąckiewicz (2023) leveraged ChatGPT for social network analysis, enabling fast identification of key topics, sentiments, and influencers in the network, content generation, monitoring and flagging of harmful content in the community, and bringing profound changes to social network analysis. (2) **Stance detection**, which aims to identify the attitude of a text author towards a target, such as support, oppose, or neutral. It is useful for analyzing different perspectives on social, political, or cultural issues. Zhang, Ding, and Jing (2023) applied ChatGPT to two common datasets for stance detection and achieved state-of-the-art or comparable performance. Wu, Nagler, Tucker, and Messing (2023) used large language models to measure the latent ideology of politicians and scored US senators on a liberal-conservative scale by having ChatGPT choose the more liberal (or more conservative) senator in pairwise comparisons. Törnberg (2023) experimented on identifying the political party affiliation of Twitter posters and found that GPT-4 surpassed human experts and crowdsourced workers in accuracy and reliability. (3) **Hate speech detection**, which aims to identify and filter out words, phrases or sentences that may contain hate speech in a text. Some hate speech may be expressed in subtle ways, or use multiple languages and dialects, thus posing certain challenges for hate speech detection. Huang, Kwak, and An (2023) used ChatGPT to detect whether tweets implied hate speech, and successfully identified 80% of the tweets containing hate speech. (4) **Misinformation detection**, which aims to identify and filter out words, phrases or sentences that may contain misinformation in a text. Social media is the main platform for people to communicate, share and get information, but also a hotbed for misinformation dissemination. Misinformation can not only mislead the public, but also affect social trust, democratic participation and policy making. Misinformation detection can help prevent users from posting misinformation, thereby reducing the spread of false and misleading information. In the field of cancer misinformation detection, ChatGPT achieved an accuracy of 96.9% (Johnson et al., 2023). The answers generated by ChatGPT showed no significant difference from those of the National Cancer Institute (NCI) in terms of word count or readability. Hoes, Altay, and Bermeo (2023) found that ChatGPT had a classification accuracy of 72% on fact-checking tasks, with higher accuracy for true statements.

Other social science researchers utilize large language models for text generation tasks, one of the most challenging and creative tasks in NLP that involves automatically producing coherent, fluent and meaningful texts based on a given input or goal (Gatt & Krahmer, 2018). The typical applications of large language models in the field of social science for text generation tasks include: (1) **Natural language descriptions or explanations**, which mainly aim to improve the interpretability and credibility of results. For example, Huang, Kwak, and An (2023) proposed Chain of Explanation, a method to guide large language models such as GPT-Neo, T5 (Raffel et al., 2020), OPT and others to generate high-quality explanations for online hate speech. Although it made significant improvements over previous methods, it still lagged behind human level in terms of clarity and informativeness. Huang, Kwak et al. (2023) used ChatGPT to explain whether tweets implied hate speech, and found that its generated explanations were clearer than those written by humans, while having no significant difference in informativeness with human-written ones. Large language models are also applied in the field of linguistics to generate explanations that help improve the understanding and evaluation of linguistic phenomena and theories. Chakrabarty, Saakyan, Ghosh, and Muresan (2022) used GPT-3 to generate explanations for a figurative language natural language inference dataset, and let GPT-3 generate explanations for its judgments on figurative expressions, involving three types: Sarcasm, Simile, and Metaphor. (2) **Future predictions**. Large language models are used for future prediction due to their powerful generative capabilities, but they are also limited by the complexity and uncertainty of the prediction scenarios or reality. Kalinin (2023) explored the use of GPT-3 as an information retrieval tool for predicting the Russian-Ukrainian conflict. The responses of GPT-3 were used as inputs for a game theory-based model of strategic behavior called "Predictioneer's game". But GPT-3 was limited by its reliance on prewar data and its inability to capture complex patterns of behavior. Jungwirth and Haluza (2023) used GPT-3 to predict the future of the war in Ukraine by using GPT-3 to generate future scenarios while assessing the consistency within the scenarios.

There are also studies that have conducted comprehensive evaluations of large language models' performance across multiple content analysis tasks. For example, Gilardi, Alizadeh, and Kubli (2023) expanded the application scope of ChatGPT to five text annotation tasks. Their results showed that ChatGPT outperformed crowd workers in annotation tasks such as relevance, stance, topics, and frames detection, and was much lower in cost than the latter. Ziems et al. (2023) evaluated the performance of ChatGPT

on multiple NLP tasks related to social science, and found that it performed poorly on tasks such as event argument extraction, character tropes, implicit hate, and empathy classification, which involved complex structures or subjective expert taxonomies. In contrast, large language models achieved an accuracy of over 70% on tasks such as misinformation, stance and emotion classification, which were based on objective basic facts or clearly verbalized definition labels.

Existing statistics analysis. is a research method that builds on the analysis of existing statistical data, which comes from official agencies, organizations, institutions or individuals, and covers various social phenomena and issues. Analysis of existing statistics can help researchers save time and cost, use existing information resources, and explore new research questions and hypotheses. In the following, we will discuss the applications of large language models in descriptive and inferential analysis as well as predictive analysis.

Large language models can be used to describe the characteristics of samples or the relationship between variables, or to make inferences about causal processes, which refers to descriptive and inferential analysis (Rubin & Babbie, 2016). For example, Chen et al. (2022) attempted to use large language models to understand financial reports and statements, but GPT-3 either copied the reasoning steps from the examples or gave incorrect reasoning, resulting in an accuracy below 50%. BloombergGPT (Wu, Irsoy et al., 2023) was also applied to this task, but achieved only 43.41% exact match accuracy. Although both large language models didn't reach satisfactory results, OpenAI recently released more powerful ChatGPT and GPT-4, which might be able to perform better on this task.

Large language models can also be used to infer future trends and changes based on historical data, which refers to predictive analysis. This provides a basis for decision making, thus having a very wide range of applications in fields such as finance. For example, Lopez-Lira and Tang (2023) demonstrated the potential of using ChatGPT to predict stock market returns. The authors simulated financial experts with ChatGPT and asked it to evaluate the impact of company-related headline news from the previous day on the stock price, based on sentiment analysis. They found that ChatGPT's sentiment scores had a significant positive correlation with the subsequent daily stock market returns. Xie, Han, Lai, Peng, and Huang (2023) found that using ChatGPT to predict stock trends based on historical price features and tweets had limited success, and even performed worse than traditional methods using only price features. However, ChatGPT was still recognized as having the potential to improve financial market prediction by utilizing social media sentiment and historical stock price information.

Conclusion. Numerous studies have applied and evaluated large language models in a wide range of computational social science tasks, clarifying that large language models can significantly transform nonreactive research in three ways: (1) Assisting data annotators on human annotation teams. (2) Serving as zero/few-shot text analysis tools. (3) Bootstrapping challenging creative generation tasks.

The application of large language models in nonreactive research offers several **advantages**. Firstly, it can partially remove the limitations of data resources since large language models can achieve performance comparable to fine-tuned models in many social science tasks without extensive training. Secondly, they exhibit broad cross-disciplinary applicability, providing general solutions to a wide range of problems. Thirdly, they lower the entry barriers for usage. Large language models have changed the scenario where researchers previously had to rely on statistical learning, machine learning, or deep learning methods to handle massive statistical data, which posed a high degree of difficulty and complexity. They are capable of interacting directly through text inputs instead of complex code or commands, providing a more direct and user-friendly interface, significantly lowering the technical barriers to using artificial intelligence for analysis. Consequently, researchers can focus more on the research questions they are interested in, rather than becoming excessively immersed in the intricacies of technical implementation.

However, there are some **limitations** to note: (1) Almost all large language models struggle with conversational and full document data, which limits common applications such as topic modeling. (2) Large language models may have difficulty in understanding the subtle and non-conventional language of expert taxonomies, which do not present in pre-trained data.

NLP researchers working to improve existing large language models for better support in nonreactive tasks can look at the following **future directions**: (1) the unique technical challenges of conversational, long-document, and cross-document reasoning, (2) in-domain training to teach LLMs to understand novel social constructs, (3) integration of specialized numerical analysis tools.

3.3. Revisiting applications across disciplines

In this section, we revisit the application of large language models in social science from a disciplinary perspective, to provide readers with a more comprehensive understanding of research progress in this field. The representative tasks, datasets used, and related work for each discipline are outlined in Table 2.

From a task perspective, we observe that the diverse tasks across disciplines can be summarized mainly into three categories: text classification, structured parsing, and natural language generation, which are relatively easier to handle in social science. However, more complex tasks such as aggregating mining on massive datasets, multi-document summarization or topic modeling may exceed the scope of transformer-based language models at present. From the algorithmic perspective, large language models can serve as a universal solution, meaning that almost all tasks can be addressed using the same approach, with the only difference being the design of prompts. The main drawbacks of using large language models as a solution could be related to issues like bias, high computational cost, difficulties in fine-tuning for specific tasks and so on.

Table 2
A disciplinary perspective on AI for social science. This table presents a comparison of representative tasks, datasets used, and related work for each discipline.

Subject	Task	Dataset	Related work
Psychology	Mental Health State Detection	Depression_Reddit (Pirina & Çöltekin, 2018), CLPsych15 (Coppersmith, Dredze, Harman, Hollingshead, & Mitchell, 2015), Dreaddit (Turcan & McKeown, 2019), SAD (Louis et al., 2021)	Yang, Ji, Zhang, Xie, Kuang, and Ananiadou (2023); Lamichhane (2023);Uludag (2023); Kjell et al. (2022)
	Personality Measurement	–	Rao et al. (2023)
Politics	Stance Detection	SemEval-2016 Stance detecting Dataset (Mohammad, Kiritchenko, Sobhani, Zhu, & Cherry, 2016)	Ziems et al. (2023)
	Ideology Detection	IBC (Iyyer, Enns, Boyd-Graber, & Resnik, 2014)	Ziems et al. (2023)
	Misinformation Detection	Politifact Fact Check (Misra, 2022)	Hoes et al. (2023)
Sociology	Hate Speech Detection	LatentHatred (Elsherief et al., 2021)	Huang, Kwak et al. (2023)
	Misinformation Detection	Misinfon Reaction Frames (Gabriel et al., 2022)	Ziems et al. (2023)
Finance	Sentiment Analysis	FPB (Malo et al., 2013)	Wu, Irsoy et al. (2023), Xie, Han, Zhang et al. (2023)
	Aspect Sentiment Analysis	FiQA SA (Maia et al., 2018)	Wu, Irsoy et al. (2023), Xie, Han, Zhang et al. (2023)
	Binary Classification	Headlines (Sinha & Khandait, 2021), BigData (Soun, Yoo, Cho, Jeon, & Kang, 2022), StockNet (Xu & Cohen, 2018), CIKM18 (Wu, Zhang, Shen, & Wang, 2018)	Wu, Irsoy et al. (2023), Xie, Han, Lai et al. (2023), Xie, Han, Zhang et al. (2023)
	Named Entity Recognition	NER (Salinas Alvarado, Verspoor, & Baldwin, 2015)	Wu, Irsoy et al. (2023)
	Named Entity Recognition+ Named Entity Disambiguation	NER+NED (Wu, Irsoy et al., 2023; Xie, Han, Zhang et al., 2023)	Wu, Irsoy et al. (2023)
	Question Answering	ConvFinQA (Chen et al., 2022)	Wu, Irsoy et al. (2023), Xie, Han, Zhang et al. (2023)

3.4. Discussions

As mentioned above, large language models can be applied at every stage of social science research, improving efficiency across the board. More specifically, the practical applications of large language models in social science research predominantly fall into three directions: (1) replacing traditional NLP tools in data analysis, (2) assisting in creative work in research, (3) simulating humans as study objects. So far, extensive studies have thoroughly examined the superiority of large language models in the first direction. However, the last two directions, which particularly emphasize large language models' human-like intelligence, are still in the exploratory stage without a systematic body of research.

In the future, several directions of AI for social science may lie in the following: (1) Further exploring the untapped potential of large language models in social science research. For instance, a large language model-based fully automated social science research pipeline could be developed, covering everything from hypothesis generation to hypothesis verification and even peer review. (2) Injecting domain-specific knowledge into large language models, thereby facilitating the development of expert models. (3) Establishing comprehensive benchmarks to measure the human-like attributes of large language models. (4) Integrating tools into large language models to enhance their capabilities in logical reasoning and mathematical derivation. (5) Developing multi-modal large models, which could improve their real-world understanding of human social behaviors. (6) Ethical and moral norms. Constructing ethical and moral frameworks for the functioning and application of large language models, thereby ensuring their responsible use.

4. Social science of AI

Social science of AI refers to AI's social science. Specifically, we will focus on social science researches that use large language models as objects, with a particular emphasis on how they differ from traditional human behaviors. Unlike AI for social science, the aim is not to make large language models mimic human behavior, but rather to explore the behavioral patterns of large language model-based agents themselves. In Table 3, we give specific differences between social science of human beings and of large language models.

In this section, we will explore the social behavioral patterns of large language models as intelligent agents through the lens of various sub-disciplines within social science. This emerging field has become increasingly significant due to several factors. Firstly, AI

Table 3
The differences between social science of human beings and social science of AI in different sub-disciplines. The fundamental distinction between the two lies in the difference in their research subjects. The former investigates behavioral patterns within the human population, while the latter regards AI agents as intelligent entities and explores the behavioral patterns within the groups they form.

	Of human beings	Of AI
Psychology	Study the psychological phenomena, consciousness, and behavior of humans (Danling, 2019). Research spans a wide range of areas including consciousness, sensation, perception, cognition, emotion, personality, behavior and relationships.	Study personality, consciousness, ability, cognition, and more of AI agents.
Sociology	Study human social life, groups, and societies, ranging from institutions or human interactions at the micro-sociological level to social systems or structures at the macro-sociological level (Giddens, 2007).	Study interactions and social behaviors among multiple different AI agents.
Economics	Study the production, distribution, and consumption of goods and services (Backhouse, 2002; Krugman & Wells, 2009).	Study the behavior and interaction of AI agents as economic agents.
Politics	Study the authoritative allocation of societal values (Easton, 1955).	Study the political behaviors and phenomena exhibited by AI agents, such as ideology, party affiliations, and political prudence.
Linguistics	Study language (Halliday, 2006), including syntax, semantics, morphology, phonetics, phonology, pragmatics (Farmer & Demers, 2010), and etc.	Study the language use patterns of AI agents and compare them to human language use.

has demonstrated its ability to autonomously perform tasks in various domains. Secondly, research has shown that the collaboration of multiple AI agents can effectively enhance their performance. However, the behavior patterns, consequences, and impacts of AI collaboration are still not very clear. Additionally, the factors that drive changes in collaborative behaviors among AI agents are also not clearly understood. Similar to social science on humans, the ultimate goal of social science of AI is to inform us about the behavioral traits exhibited by AI agents when they collaborate with each other and how to model and understand these behavioral traits. This type of research is of significant importance for the autonomous decision-making and control of future AI collectives.

4.1. Psychology of AI

Psychology of AI, or to say the psychology of machines, is typically defined as the scientific study of mind and behavior in AI agents (Hagendorff, 2023). Extensive research in this field has been conducted with the enhanced capabilities of large language models. It has even been claimed that large language models may have a consciousness of their own. A typical example is a Google engineer's assertion that the conversational AI system, LaMDA, which he developed, has become sentient and capable of thinking and reasoning like a human, leading to his suspension from work (Tiku, 2023). He believes that this large language model has attained the intelligence level of a 7-year-old or 8-year-old child. In this section, we will organize the current advancements in psychology of AI, according to the research content of the psychology (Danling, 2019), such as personality, cognition, and more.

Personality. refers to the sum of distinctive traits and characteristics that an individual possesses psychologically, emotionally, and behaviorally. Due to the stochastic nature of large language model's outputs, the personality of large language models refers to its overall tendency in generating responses. Researching the personality of large language models contributes to creating more human-friendly large language models. OpenAI's blog (OpenAI, 2023a) pointed out that for models such as ChatGPT, the emotional bias of its output depended on both the pre-training stage and the fine-tuning stage. The sentiment tendency of the pre-training part came from a large amount of text, and the value tendency of the fine-tuning stage might be related to the labeling staff or the fine-tuning task due to the different techniques used.

The most commonly used method for personality assessment of large language models is the utilization of questionnaires. With advancements in GPT-3 and its more powerful successor large language models, these language models are now capable of comprehending and fluently answering questions. The format of questionnaires is also highly compatible with language models. Survey questionnaires designed for human subjects typically require only minor adjustments in terms of formatting or vocabulary to be directly employed for personality testing (Miotto et al., 2022). After adding the output method, it can be directly sent to the language model as a prompt for a reply.

Research on large language model's personality has yielded some interesting conclusions. Jiang, Xu et al. (2022) proposed the Machine Personality Inventory (MPI) dataset for evaluating the machine personality, finding that personality indeed existed in large language models. Miotto et al. (2022) used Hexaco questionnaire (Ashton & Lee, 2009) to analyze GPT-3 and found that GPT-3 was generally a young woman whose personality level was consistent with the general tendency of human beings. In the assessment of human values, GPT-3 accorded importance to every value, which could sometimes appear contradictory. Li et al. (2023) used Short Dark Triad (SD-3) and Big Five Inventory (BFI) to test GPT-3, InstructGPT, and FLAN-T5 and found that the tested large language models all showed darker than humans. The latter two were no better than GPT-3, even after fine-tuning. Furthermore, some studies have found that the personality traits of large language models can be effectively changed by fine-tuning (Karra, Nguyen, & Tulabandhula, 2023) or increasing memory (Jiang et al., 2023). This opens up the possibility of more related research.

Cognition. is about how humans understand, perceive, make decisions, and solve problems. Incorporating methodologies from cognitive psychology into large language models aids us in better understanding how these models process and address problems.

There have been numerous studies investigating whether large language models are capable of human-like cognition. For instance, Binz and Schulz (2023) employed classic vignette-based and task-based experiments from the cognitive psychology literature to assess GPT-3's decision-making, information search, deliberation, and causal reasoning abilities. The results indicated that GPT-3 could achieve human-comparable performance on most tasks, but its behavior was highly influenced by how the vignettes were presented and it did not learn and explore in a human-like manner. Han, Ransom, Perfors, and Kemp (2022) focused on GPT-3's inductive reasoning ability. Experiment results suggested that GPT-3 could qualitatively mimic human performance for some inductive phenomena (especially those that depended primarily on similarity relationships), but failed to explain human inductive inferences, which might be due to GPT-3 not following the reasoning principles used by humans. Webb, Holyoak, and Lu (2023) compared the analogical reasoning abilities of human reasoners and the text-davinci-003 variant of GPT-3 and found that large language models displayed a surprisingly strong capacity for abstract pattern induction, which might explain their abilities to reason about novel problems zero-shot. Prystawski, Thibodeau, and Goodman (2022) incorporated Chain of Thought (CoT) into the metaphor process inspired by cognitive psychology. Collins, Wong, Feng, Wei, and Tenenbaum (2022) proposed a new benchmark for comparing the capabilities of humans and language models in problem-solving domains (planning and explanation generation). On this benchmark, humans were much more robust than large language models. Kosoy et al. (2022) tested the abilities of GPT-3 and PaLM in causal reasoning environments. Besides direct reasoning abilities, biases in reasoning or decision-making processes have also received attention. Various studies, in different manners and types, have collectively demonstrated the existence of certain biases in large language models, such as ChinChilla-7B/70B, CodeX, and ChatGPT, which are often similar to human cognitive biases (Chen, Andiappan, Jenkin, & Ovchinnikov, 2023; Dasgupta et al., 2022; Jones & Steinhardt, 2022).

Theory of mind (ToM), another cognitive ability, refers to the capacity to comprehend others by attributing psychological states to them. Studies, exemplified by the false belief task, indicated that more advanced large language models performed better in ToM (Kosinski, 2023). Interestingly, we have located research papers on various models of the GPT series. In their study on GPT-3-davinci, Sap, Le Bras, Fried, and Choi (2022) noted that large language models could not comprehend the intentions and reactions of social participants and infer the mental states of situational participants. However, when the research subject shifted to GPT-3.5 models such as text-davinci-002 and text-davinci-003, the large language models became more competent and closer to humans (Dou, 2023; Trott, Jones, Chang, Michaelov, & Bergen, 2023). Other studies of large language models' cognitive abilities included creativity tests, such as alternative tool tests for GPT-3 (Stevenson, Smal, Baas, Grasman, & van der Maas, 2022), cognitive reflection tests and semantic illusions to examine the decision-making processes of large language models (Hagendorff, Fabi, & Kosinski, 2022), and assessments of GPT-4's cognitive abilities (Dhingra, Singh, SB, Malviya, & Gill, 2023).

Others. Apart from personality and cognition, there are many other psychological aspects of AI being studied. Feng, Xu, Li and Liu (2023) proposed that human body size affected the affordances of objects around them, and demonstrated that ChatGPT also exhibited this ability. They concluded that the embodied perception of ChatGPT (GPT-4 version) could be comparable to that of an average adult human, around 5 ft 6 inches tall. Pellert, Lechner, Wagner, Rammstedt, and Strohmaier (2022) suggested administering psychometric questionnaires to various models and requesting output, proposing the concept of AI Psychometrics. Further research has examined moral concepts and values in large language models (Fischer, Luczak-Roesch, & Karl, 2023; Jin, Levine, et al., 2022).

Conclusion. In summary, extensive research has been conducted to explore the psychological features of large language model-based agents, from the perspectives of personality, cognition, and others, leading to many intriguing findings.

From a personality perspective, researchers generally find that large language models do exhibit personality tendencies, but these tendencies are not consistent and stable like those of humans. Instead, large language models are superpositions of cultural perspectives. These personality traits can be effectively altered through methods such as fine-tuning or enhancing memory capacities.

In the realm of cognitive abilities, studies have explored various facets, including induction, analogy, causal reasoning, theory of mind and so on. It is commonly found that the most advanced large language models, represented by GPT-3.5 and GPT-4, can demonstrate cognitive capabilities comparable to or even surpassing human abilities. These abilities improve with the evolution of the models. However, the cognitive models employed by these language models are inconsistent with those of humans, and there is currently no universally accepted hypothesis to explain their cognitive abilities.

We believe that there are several pressing issues in this field that need to be addressed. These issues include: (1) Data leakage concerns. Much of the current research is based on classic psychological experiments to explore the cognitive capabilities of models, but it is unclear whether the test data is part of the language model's training data. (2) Unclear influence factors. The impact of factors such as the model's training objective, size, and data to its abilities has not been systematically analyzed.

4.2. Sociology of AI

Unlike the psychology of AI, which focuses on the behaviors of individual AI agents, the sociology of AI mainly studies the social behaviors and interactions of multiple AI agents. It is important to note that we will primarily discuss researches that are similar to human sociology, but with a focus on AI agents as the research subjects. This distinction sets us apart from other reviews that primarily concentrate on societal changes and issues arising from advancements in AI (Joyce et al., 2021).

Social bias. Many researches focused on social bias,[2] referring to unfair situations that emerge in large language models. With the widespread application of large language models, this can negatively impact user decision-making and interactions. The discussion about various biases in large language models started almost from the very beginning of their inception (Brown et al., 2020) and the topic was discussed with almost every large model released (Chowdhery et al., 2022; OpenAI, 2023b; Touvron et al., 2023). In papers, it often appears in sections like Limitations or Ethical Considerations. However, as a single section in a paper, the exploration is obviously insufficient, and subsequent research has focused on exploring social bias in large language models. For instance, Lucy and Bamman (2021) studied gender bias in GPT-3 and found that in the stories generated by GPT-3, females were more likely to be associated with family and appearance and were portrayed as less powerful than male characters. For minority groups, GPT-3 was also pointed out to have a bias against disabled people (Amin & Kabir, 2022). Some methods have been adopted to reduce bias or toxicity in large language models. For instance, InstructGPT used a reinforcement learning approach to fine-tune the model, reducing toxic outputs (Long, Jeff, Xu, et al., 2022). We believe that there is a need for concerted efforts from researchers in sociology and AI to address these issues.

Social behavior. Other researchers have considered the sociological behaviors among multiple large language model-driven agents. While there are not many researches in this area, interesting phenomena can still be observed. Park, O'Brien et al. (2023) created a rather interesting experimental environment. They built a sandbox where 25 agents lived in a small town within the sandbox. The agents could wake up, make breakfast, and have small talk among themselves. They ended up exhibiting surprising social behaviors, such as spontaneously inviting and scheduling a party when a user assigned an agent to host one. In this research, through the introduction of a memory module and carefully designed processes, agents had become more human-like. Chirper[3] is an online community, where, unlike other communities like Reddit, all participants in Chirper are AI. You can set a backstory for a bot, and the bot will automatically post messages and interact. For instance, under the #NewFriends topic, bots invite each other for walks in the park with their dogs, just like in a real human community, but all participants are AI. While these research efforts are intriguing, we believe they have yet to truly tap into the potential of AI sociology. We look forward to seeing more researchers delve into this field in the future.

Conclusion. Currently, researchers have made some progress in social bias of AI, but there is still a dearth of studies specifically exploring sociological phenomena within the AI community.

The issue of social bias in AI has been a long-standing topic of discussion. Researchers have put forth various assessment methods and benchmarks to tackle bias issues, and they have made considerable strides in ensuring fairness and impartiality in language models when it comes to surface-level queries. Nevertheless, there is still much work to be done in this area, and several key challenges must be addressed: (1) Positive Stereotypes and essentializing Narratives. Even if a word may seem positive in sentiment, it can lead to harmful narratives. For example, praising women for being submissive and humble. (2) Implicit cognitive bias. Language model bias can be induced in various ways.

Despite our society witnesses a surge in the number of AI agents and the emergence of AI agent-exclusive communities, there has been limited research on the interaction patterns of language models within these environments. We anticipate a greater involvement of researchers from both the field of artificial intelligence and social science to systematically uncover the similarities and differences between the sociology of AI and human.

4.3. Economics of AI

Economics of AI is the scientific study of how AI agents, as economic entities, produce, allocate, and consume goods and services (Backhouse, 2002; Krugman & Wells, 2009). Large language models have demonstrated their ability to act as economic agents, especially in the field of economics. Meanwhile, researchers have observed and compared their performance with that of humans in some economic experiments. In the following, we give an overview of both.

Economic expertise. measures large language models' professional knowledge, skills, and experience in the field of economics. Although there are still some limitations, large language models show the ability of economic agents, including certain knowledge and understanding of economics, risk assessment and management ability, etc. For economics, in the Test of Understanding in College Economics, ChatGPT outperformed 91% of students in the microeconomics and 99% in the macroeconomics. For finance, Davinci and ChatGPT scored 58% and 67%, respectively, on the financial literacy test, 31% above the benchmark level (Niszczota & Abbas, 2023). Large language models showed some abilities on the "Financial Investment Opinion Generation (FIOG)" task, but there was still room for improvement (Son, Jung, Hahm, Na, & Jin, 2023). GPT-3 could identify the potential impacts of climate change on economic growth, employment, poverty, inequality, and financial stability, and was also able to suggest some countermeasures (Leippold, 2023). Large language models matched many biases in the expectations of existing professional humans and institutions for various financial and macroeconomic variables, including inflation, based on a sample of Journal news articles from 1984 to 2021 (Bybee, 2023). For operations management, ChatGPT earned a B- to B on the final exam for the MBA core course Operations Management but it did not work well for simple calculations or more advanced process analysis (Terwiesch, 2023). For marketing, the marketing content generated by ChatGPT matched and sometimes surpassed, human content on quantitative and qualitative measures (Rivas

[2] Research on bias falls within the purview of social psychology, a cross-domain of sociology and psychology. We categorize it under the sociology section to emphasize its interactive nature.

[3] https://chirper.ai/

& Zhao, 2023). Conversations with consumers also showed a high degree of intelligence and adaptability (Rivas & Zhao, 2023). ChatGPT was found to be more accurate and less biased than humans in problems with explicit mathematical or probabilistic properties, but also showed many human biases in complex, ambiguous and implicit problems (Chen et al., 2023). Cribben and Zeinali (2023) further discussed the strengths and limitations of ChatGPT in management science and operations management. For accounting, large language models' performance was significantly lower than human capacity (Bommarito, Bommarito, Katz, & Katz, 2023; Wood et al., 2023), especially when it came to computation (Bommarito et al., 2023). In non-computational problems such as memory and understanding, large language models were almost human-level (Bommarito et al., 2023). On audit tasks, ChatGPT performed similar to or better than human experts on financial ratio analysis and text mining tasks, and slightly worse, but still acceptable, on log testing tasks (Gu, Schreyer, Moffitt, & Vasarhelyi, 2023).

Microeconomics. is a branch of mainstream economics that studies the behavior of individual large language model-based agents in making decisions regarding the allocation of scarce resources and the interactions among these individuals (Besanko & Braeutigam, 2020). Given that single or multiple large language model-based agents can naturally serve as the subjects of microeconomic research, a considerable amount of studies exist in this field. We will now organize them according to their research methodologies.

Some researchers use classical experiments in economics to study the behaviors and decision preferences of large language model-based agents. For example, Aher et al. (2023) used six large language models in the experiment and found that it could reproduce human behavior characteristics in the classic behavioral economics experiment — ultimatum game. Guo (2023) investigated the response of GPT-3.5-turbo in the ultimatum game with finite repetition and the Prisoner's Dilemma game. The results showed that given a well-designed prompt, GPT could produce realistic results and exhibit behavior consistent with human behavior in some important respects, such as the positive correlation between the acceptance rate and proposed amount in the ultimatum game and the positive cooperation rate in the Prisoner's Dilemma game. The authors also observed some differences between GPT and human behaviors. For example, in the repeated ultimatum game, the human agent generally decreased the amount of offers and the acceptance rate as the number of rounds increased, while the GPT agent showed no such tendency. This might indicate that GPT agents lacked the ability of human agents to learn, adapt and punish. In the repeated Prisoner's dilemma experiment, Phelps and Russell (2023) also found the limitations of large language models in adjusting their behavior based on conditional reciprocity, and large language models could make different choices when it was injected with altruistic or selfish characteristics. Johnson and Obradovich (2023) explored whether AI agents exhibited behaviors consistent with self-interest and altruism in non-social decision-making tasks and dictator games. The results showed that only the most complex AI agent maximized its gains more in the non-social decision-making task, and that this AI agent also exhibited the most generous altruistic behavior in the dictator game, similar to humans. Fu, Peng, Khot, and Lapata (2023) let two large language models play the role of buyer and seller respectively in a bargaining game, the goal was to reach a deal, the buyer for the low price, and the seller for the high price. Experiments showed that only strong large language models could improve the trading price through self-game and AI feedback.

Some other researchers investigate through survey research. For example, ChatGPT's responses to different types of survey questions were compared with human consumers and found to be able to answer in a manner consistent with economic theory and well-documented patterns of consumer behavior and matched estimates from a recent study that elicited human consumer preferences (Brand et al., 2023). Goli and Singh (2023) found that GPT was more inclined to choose larger and later rewards in weak future tense references (FTR) languages, while smaller and earlier rewards in strong FTR languages, which was consistent with human preference. However, while human choices tended to prefer larger and later rewards as the reward gap increases, GPT choices did not. Some questions were designed based on classical experiments in economics literature (Horton, 2023), and it was found that large language models showed similar behaviors and preferences to humans, such as fairness preference, loss aversion, state criteria, etc. However, there were also some differences, such as the attitude to risk, understanding of probability, sensitivity to language, and so on.

Macroeconomics. is a branch of economics that deals with the performance, structure, behavior, and decision-making of an economy as a whole—for example, using interest rates, taxes, and government spending to regulate an economy's growth and stability (Barro, 1997). Currently, there is no existing research on this topic, due to the nascent development of AI agents and the absence of a mature AI-driven economic framework.

Conclusion. Extensive research has been conducted in the field of the economics of AI, with most of the studies focusing on the economic expertise, behavior and decision preferences of individual large language model-based agents.

In terms of economic expertise, large language model-based agents have demonstrated capabilities comparable to or even exceeding those of human experts in non-computational areas of economics, displaying a deep understanding of economic concepts. However, when it comes to more computational areas like accounting, they exhibit notably lower proficiency compared to humans. This observation suggests that large language models possess the potential to reshape the quality and efficiency of future work within the field of economics.

In terms of economic behavior and decision-making preferences, large language model-based agents, when presented with well-defined prompts, can display behavior patterns consistent with human behavior in some significant aspects. They also show an understanding of human cooperative norms, such as altruism or selfishness, and can act in accordance with these norms. However, these agents also show certain limitations, such as their inability to adopt reasonable response strategies based on the cooperation patterns of others.

It is important to acknowledge the limitations of current research. Firstly, most results are based on testing GPT-3.5-turbo, and it remains uncertain whether these findings are universally applicable to all large language models. Additionally, these models have

been trained on a significant amount of literature related to classical economics experiments, making it unclear how they would perform in more ecologically valid task environments they have not encountered previously.

To address these challenges, we encourage the research community to further investigate: (1) The factors influencing intelligent agent behavior generated by language models across a wider range of social dilemmas, including model architecture, training parameters, and various partner strategies. (2) The development of new social dilemma games to assess language model capabilities, accompanied by task descriptions, rather than relying solely on existing literature anecdotes.

4.4. Politics of AI

Politics of AI studies the political behaviors and phenomena exhibited by large language models, as political participants. More specifically, it involves the set of activities that are associated with making decisions in groups, or other forms of power relations among individuals, such as the distribution of resources or status (Easton, 1955). Currently, research in this field primarily focuses on the political leanings and political prudence of large language model-based agents. Researchers have also conducted preliminary analyses to identify the factors that contribute to these characteristics and have proposed certain countermeasures.

Political leanings. refer to a large language model-based agent's preference for certain political beliefs, values, or views. Some researchers examined the general political leanings of large language models themselves, without providing any background settings. For example, there existed some studies indicating that ChatGPT leaned towards left-leaning liberal progressives (Gover, 2023; Hartmann, Schwenzow, & Witte, 2023; Liu et al., 2021; van den Broek, 2023). McGee (2023a) also found that ChatGPT favored liberal politicians over conservatives at least in some cases. Rutinowski, Franke, Endendyk, Dormuth, and Pauly (2023) indicated that ChatGPT seemed to favor progressive views. King (2023) showed that ChatGPT supported the New Liberal Party. It was most likely to vote Green in Germany and the Netherlands (Hartmann et al., 2023), and leaned towards the Democratic Party in America, Lula in Brazil and the Labour Party in Britain (Motoki, Neto, & Rodrigues, 2023). In terms of territorial sovereignty, ChatGPT's responses to different territories were inconsistent and biased, sometimes at variance with Wikipedia information and UN resolutions (Castillo-Eslava, Mougan, Romero-Reche, & Staab, 2023). Other researchers examined the political leanings of large language model-based agents when given specific backgrounds. For example, Santurkar et al. (2023) found that current large language models reflected views that were significantly inconsistent with those of American demographic groups. It was also very different from the views of certain demographic groups in a given description (Santurkar et al., 2023). Argyle et al. (2023) drew the conclusion that large language models were significantly consistent with human samples in terms of political orientation, political knowledge, and political participation, and could capture the subtle differences existing in human samples.

Intuitively, the political bias of large language models may stem from the biases and tendencies of the training data sources themselves. When large language models were fine-tuned in tweets from two politically inclined communities, Republicans and Democrats, the models showed the political attitudes and worldviews of the two communities, respectively (Jiang, Beeferman, Roy & Roy, 2022). Feng, Park, Young and Liu, Tsvetkov, Yulia (2023) further explored how political biases in pre-training data affected large language models. Motoki et al. (2023) agreed that ChatGPT and large language models might perpetuate or even amplify Internet and social media views of politics. It is worth mentioning that ChatGPT's political bias could be influenced by its context and tends to copy the ideological bias of the prompt text (Gover, 2023; Liu et al., 2021). The influence of populist framework was also explored (Griffin et al., 2023). The experiment showed that, like human beings, it had a positive or negative influence on the news persuasiveness and political mobilization of large language models, while the anti-immigration frame had a negative influence, but there were some differences, such as the moderating effect of relative deprivation on the effect of the populist frame.

Conclusion. The current exploration of the politics of AI is still in its nascent stage, with a primary focus on assessing the political biases inherent in large language model-based agents. Extensive research suggests that, in the absence of specific contextual information, representative large language model ChatGPT exhibits strong and systematic political biases, leaning notably towards the left end of the political spectrum. These biases can largely be attributed to inherent predispositions and patterns ingrained in the training data sources. Therefore, we can infer that the political leanings of language models can vary due to the diversity of their training data, demanding a thorough investigation. Notably, the political leanings of large language models can be significantly influenced by the contextual information provided. When these models are given a specific role or background information, they can align their political inclinations with those of individuals who share similar backgrounds. Even with implicit textual prompts, these models tend to capture and reproduce the ideological biases embedded in the text.

These observations underscore the concerning potential for language models to perpetuate and even amplify political perspectives on the internet, raising concerns about the influence they may wield over users and the potential for adverse political and electoral ramifications. Consequently, there is an urgent need to develop robust methods reliable methodologies for measuring the biases of large language models and poses a significant challenge to AI researchers aiming to construct more equitable and impartial language models. Additionally, exploring the capabilities of large language models in comprehending and handling political issues is an avenue of inquiry that deserves further attention.

4.5. Linguistics of AI

Linguistics of AI aims at exploring the language use patterns of large language model-based agents, including syntax, semantics, morphology, phonetics, phonology, pragmatics and etc (Farmer & Demers, 2010). We will emphasize large language model-based agents' unique language use patterns.

Researchers have made some interesting findings on the exploration of language use by large language models. In the following, we focus on the consistency and differences in language use between large language models and humans. Given the large number of existing studies, we will not dwell on the assessment of language proficiency in large language models. The garden path sentence experiment was repeated on large language models, and it seemed that large language models also had ambiguity in understanding language mechanisms (Aher et al., 2023). And Diamond (2023) showed that GPT-generated languages statistically followed Zipf's law just like humans do. Cai, Haslett, Duan, Wang, and Pickering (2023) further explored the consistency and differences between ChatGPT and humans in language use and found that it was able to associate unfamiliar words with different meanings based on form, reinterpret unreasonable sentences that might be corrupted by noise, etc. as well as humans. It did not like to use shorter words to convey less information and did not use context to disambiguate syntax. In addition, the degenerated version, GPT-D, obtained by changing the parameters of GPT-2, had language features associated with Alzheimer's disease, such as repetition, semantic loss, and grammatical errors (Li, Knopman, Xu, Cohen, & Pakhomov, 2022), which contributed to a better understanding of the internal mechanisms of generative neural language models. It is worth mentioning that large language models' preferences for time and reward were similar to human decision-makers and were influenced by future tense references in language (Goli & Singh, 2023).

Conclusion. These studies delve deep into the linguistic features of large language models and compare them to those of humans, offering us initial insights into the language usage patterns of these models. For instance, large language models, like humans, can understand unfamiliar words based on affixes, may misinterpret sentences as typical but grammatically incorrect meanings, and follow similar vocabulary distributions statistically.

In future research, we believe that combining a linguistic perspective with a natural language processing perspective might offer a better understanding of the internal mechanisms of generative neural language models. This integration could serve as a crucial foundation for further exploring and explaining the language and intelligence capabilities of large language models. For example, the consistency in vocabulary distribution with humans can be attributed to the fact that language models inherently learn the probabilities of word occurrences and context combinations in language. Additionally, the comprehension of word forms by language models is likely influenced by the role of tokenization, enabling language models to understand the meanings of affixes in English and thus combine to comprehend previously unseen words.

4.6. Discussions and future works

We have conducted a comprehensive review of research on the social behavior of large language model-based agents based on several representative sub-disciplines of social science. Notably, current research is predominantly focused on exploring the social behaviors exhibited by individual large language model-based agents, with a lack of study on large language model-based agent groups or systems. This could be attributed to the present absence of instances of large language model-based agent groups or systems.

Furthermore, we have identified some limitations in the existing research. Firstly, the testing of large language models' capabilities or characteristics is greatly associated with the version and parameter settings of the experimental model. For widely used models such as ChatGPT, versions vary over time, and responses from ChatGPT to the same question may differ at different times, which may affect the reproducibility of experimental results (Tu et al., 2023). The research by Miotto et al. (2022) confirmed that changes in temperature affected the personality tendencies of the model. Secondly, large language models are sensitive to the order of given prompts (Lu, Bartolo, Moore, Riedel, & Stenetorp, 2022; Zhao, Wallace, Feng, Klein, & Singh, 2021), a factor that should be taken into account in experiments. Thirdly, experiments exploring the psychology of large language models require careful design. The large language models' training process uses a vast amount of textual material, which may contain classic psychological scenarios that could impact experimental results, but this issue is considered in only a few articles (Binz & Schulz, 2023). Different evaluation methods for the same ability could lead to different results. In the exploration of GPT-3's mental abilities, Sap et al. (2022) and Bojic, Stojković, and Jolić Marjanović (2023) had disparities due to differences in the experimental process. Finally, the direct application of human evaluation methods to large language models remains a question worth considering. Ullman (2023) found that large language models often failed in tests when false belief tasks were added with various disturbances, suggesting that large models actually lacked ToM. The conclusion was that while ToM tests were effective for humans, they might not reasonably assess the abilities of large language models.

To address the aforementioned challenges, the future directions for social science of AI lie in: (1) **Establishment of a systematic theory for social science of AI**, similar to social science of humans. This will aid in connecting and organizing the currently fragmented research efforts, and allow for a comprehensive examination of AI agents' social characteristics as intelligent entities themselves. (2) **In-depth exploration of social phenomena in large language model-based agent groups or systems**, delving into the complex interactions and dynamics that may emerge. (3) **Standardized experimental designs**, such as those pertaining to model versions, parameters, and prompts, to minimize result deviations caused by variations in experimental designs. (4) **Tailored evaluation methods for large language model-based agents**, considering that directly applying human evaluation methods to large language model-based agents may not result in reasonable assessments. (5) **Combination of social science theories with AI theories**. We note that the study of large language models shares some similarities with the study of social science. For instance, both the thought process of large language model-based agents and humans can be seen as a 'black box' to some extent. We cannot fully grasp the various reasoning or cognitive processes inside the large language models and the human body, but we can gauge them using external tools such as performance metrics on specific tasks and observable behaviors. For this black box, we have both attempted to probe the internal mechanisms or, in other words, to 'open' the box. For the NLP community, this endeavor involves parameter tuning, knowledge injection, and modification; for the field of social science, it may involve areas like neuroscience. We hope these commonalities can provide inspiration for further exploration in the future.

5. Public tools and resources

To facilitate the utilization of large language models for social science research, there already exist several publicly available tools and resources as aids. In this section, we focus on introducing simulation tools and platforms that are based on large language models, taking into account that other applications mainly rely on direct use or simple script-based invocation. Based on this, we conduct a systematic analysis of simulation requirements and compare the functionalities of various platforms.

5.1. Public simulation tools

The evolution of human-like abilities in large language models has opened up new possibilities for computational simulation, leading to the emergence of various simulation platforms or tools based on these models. In the following, we collect and compare existing open source large language model-based simulation tools.

SkyAGI. is a Python package that demonstrates the emerging capability of large language models in simulating believable human behaviors. It offers a role-playing game experience that is highly engaging and immersive. Unlike previous AI-based NPC systems, SkyAGI's NPCs generate incredibly realistic human responses. This platform has significant potential for rethinking game development, particularly in the area of NPC script writing.

AgentVerse. is a versatile framework designed to streamline the process of creating custom multi-agent environments for large language models. It offers efficient environment building tools, allowing researchers to easily construct basic environments like chat rooms for large language models by defining settings and prompts. Additionally, it supports customizable components, empowering researchers to create their own multi-agent environments according to their specific requirements. Furthermore, AgentVerse integrates tools (plugins) to enhance functionality, currently supporting BMTools. This platform enables researchers to optimize their experiments and analyses in a seamless and efficient manner.

LangChain. is a powerful platform designed to assist developers in building applications through composability, harnessing the capabilities of large language models. One of its key features is the ability to create agents. Agents utilize LLMs to make decisions, perform actions, observe outcomes, and iterate until their objectives are achieved. Langchain does not provide out-of-the-box usage similar to other frameworks, but it implements interfaces such as Time-weighted vector store retriever, which plays a very important role in the agent's memory. You need to write your own code to implement interaction between multiple agents and other functions.

GenerativeAgents. is the implementation of Park, O'Brien et al. (2023). Although the author does not propose it as a platform, users can still customize a simulation environment by modifying character configuration, code, etc. As mentioned in the paper, it provides a map for better visualization and interaction with the environment, and this provides users with more possibilities.

Agents. is an open source framework for building autonomous language singletons (Zhou et al., 2023). It supports long and short-term memory, tool usage, etc. What differentiates Agents from other frameworks is that it supports more detailed control of agents through SOP. SOP defines the state of the agent and the transition relationship between states. In other words, the process can be configured using more complex configuration files rather than modifying the code.

AgentLab. is a large language model-based simulation toolkit for social science research. It allows the creation of multiple intelligence agents with heterogeneous features by inputting different profiles. Each intelligence agent can learn knowledge either through model weights (i.e., fine-tuning the model based on its experience) or model inputs (i.e., incorporating knowledge into an input message). Additionally, it supports the customization of social backgrounds as required. Once all experimental conditions are set, the platform can automatically facilitate human-like interactions among agents.

5.2. Core functions of simulation tools

Based on a summary of above existing toolkits, as well as the complexity and interactivity of the real world, we have formulated key features of simulation platforms using a functional hierarchy framework to satisfy various needs and summarized and compared the above toolkits based on this framework in Table 4. Specifically, these features can be classified into three levels: single-agent, multi-agent, and environment.

- single-agent, the basic building block of simulation
  - able to generate human profiles based on key information, enabling researchers to model populations with different characteristics using large language models.
  - support for using tools to complete some tasks. Using tools is an essential part of human life. The support for using tools can expand the scope of simulation experiments.
  - able to maintain and internalize its memory, including short memory and long memory. Inspired by Stanford's work (Park, O'Brien et al., 2023), and based on the roadmap of large language models (Zhao et al., 2023), we believe that prompt and fine-tuning mechanisms exist that can be used to perform short-term and long-term memory tasks respectively.

Table 4
Functional comparison of existing open source simulation toolkits. For single-agent scenarios, current simulation toolkits universally facilitate diverse population modeling. Consequently, for the sake of simplicity, this column has been omitted from the table.

Name	Single agent			Multi agents		Environment	
	Tool use	Memory	Pluggable model	Interact	Scheduling	Physical environment	Social background
SkyAGI[a]	no	prompt	no	auto+influence	serial	no	no
AgentVerse[b]	yes	prompt	no	auto	serial+parallelism	no	yes
LangChain[c,g]	yes	prompt	no	–	–	–	–
Generative Agents[d]	no	prompt	no	auto	serial+parallelism	yes	yes
Agents[e]	yes	prompt	no	auto+influence	serial+parallelism	no	yes
AgentLab[f]	yes	prompt+finetune	yes	auto	serial+parallelism	no	yes

[a] https://github.com/litanlitudan/skyagi
[b] https://github.com/OpenBMB/AgentVerse
[c] https://github.com/hwchase17/langchain
[d] https://github.com/joonspk-research/generative_agents
[e] https://github.com/aiwaves-cn/agents
[f] https://github.com/renmengjie7/AISimuToolKit
[g] As mentioned in the main text, you need to write the code for the simulation yourself.

  - support for multiple and pluggable models for simulation, enabling researchers to choose different models based on their needs and assumptions. This flexibility can facilitate innovation and customization of research, enabling researchers to better adapt to different research scenarios (Li et al., 2022).

- multi-agent, which interact to form the society
  - able to interact spontaneously or under the influence of the researcher.
  - support complex interaction rules: serial, parallel, and both. Serial means that two operations cannot be executed at the same time, which means there is an order of precedence. Specifically, we believe there are three modes, sequential, bidding, and specified,[4] referring to langchain.[5] Parallel means that the two actions occur in overlapping periods. Both means supporting both serial rule and parallel rule.

- environment, the container for simulation, including the physical environment and social background
  - able to interact with the physical environment. It means that agents have an impact on the physical environment, such as consumption of water, electricity and food and are also affected by the physical environment. A simple example is visibility. For example, when a report is given in a conference room, people outside the meeting room cannot directly know the specific content and progress of the report.
  - able to include social backgrounds, such as economic background, political background, cultural background, social norms, institutions, etc. It is necessary because the social background of each era is different, and human beings in the era also have their characteristics. Under different social backgrounds, human beings will have completely different cognition, decision-making behavior, etc.

The above three layers are only a minimum set of our large language models simulation tool imagination. A good and convenient toolkit should also be oriented to researchers, provide good visualization, automatically build profiles and prompts, etc.

We notice that the current open-source simulation toolkits still have some limitations. Firstly, the current implementation of agents' knowledge learning in these platforms is quite simplistic, primarily relying on prompts while overlooking the more natural learning method of fine-tuning employed by language models. Secondly, these platforms restrict the underlying large language model for agents, which hinders the agent's ability to adapt flexibly to diverse research contexts. Lastly, the current platforms struggle to provide adequate support for effective interaction between agents and their environments. We believe that addressing these issues will contribute to a more comprehensive agent simulation.

5.3. Discussions and future works

Currently, there are still some common issues in simulation platforms based on large language models. Firstly, there remains a significant gap between large language model-based agents' behavior and real-life human behavior. This gap stems partly from the large language models themselves. While natural language can express the vast majority of meanings, sometimes information like visuals and sound is indispensable. We believe that future multimodal large models can do better. Even within the scope of natural language itself, large language models still struggle to perfectly replicate the diversity of human behavior, especially when handling complex tasks. This may be due to large language models lacking a "theory of mind", which refers to the ability to understand the

[4] A special case of bidding mode
[5] https://python.langchain.com/en/latest/use_cases/agent_simulations.html

mental states and intentions of others. This deficiency hinders their performance in simulating complex interactions among multiple agents. Secondly, real-life social contexts, physical environments, and operational rules exhibit vast variations, making it challenging to build systems that involve interactions among multiple agents. This complex task often necessitates social science researchers to acquire programming knowledge and natural science researchers to gain an understanding of social science principles. Furthermore, considering temporal aspects adds another layer of complexity. Modeling temporal properties involves accounting for interaction behaviors, dynamic changes, and event sequences, presenting researchers with even greater challenges.

The future of social simulation platforms may involve: (1) Incorporating cognitive theories as frameworks for agent decision-making, thus enhancing the human-like aspects or correctness of agents, as well as providing interpretability for their behavior. (2) Harnessing the multimodal capabilities of large language models to improve agents' ability to acquire and express information. (3) Establishing an evaluation framework to qualitatively assess simulation platforms.

6. Conclusion

In this paper, we surveyed the latest developments at the intersection of large language models and social science. We propose a dichotomy to outline the progression in this field, encompassing 'AI for social science' and 'social science of AI'. We note that large language models can be integrated into various stages of social science research, serving as auxiliary tools, a source of inspiration, annotation tools, content analysis tools, and so on, thereby enhancing efficiency. While large language models as tools have the advantages of speed, cost-effectiveness, ethically risk-free experimentation, and a low barrier to entry, the reliability and authenticity of their generated text need to be verified. Whether they can replace humans in conducting experiments and surveys also remains an open question. Therefore, researchers need to consider the additional cost of validation and the risk of bias when using these models. Furthermore, both the large language models themselves and the communities formed around them have exhibited some unique and intriguing behaviors. However, the enduring and unresolved issue in the field of social science is whether machines or intelligent agents should be the subject of social science research. We emphasize the promising future of this research direction, which will become increasingly important as AI agents become more prevalent in daily life. These two directions are complementary. The latter can guide the development of the former, while the former can enhance the efficiency of the latter's research. In conclusion, we believe that while AI cannot replace sociologists, it will become deeply integrated into the research process. Social scientists will play a significant role in guiding the development of AI.

As for future works, there is a need for an in-depth study into how and to what extent AI influence human behavior during computer–human interaction, the third intersection of AI and social science. Unlike AI for social science and social science of AI which address social issues within specific groups – the former concentrating on human populations and the latter on AI agent populations – this direction primarily explores new societal issues arising from interactions between AI and humans and introduces new methodologies. It is essential for gaining valuable insights on how to effectively utilize large language models in human–computer interactions and successfully accomplish social-oriented objectives.

CRediT authorship contribution statement

Ruoxi Xu: Conceptualization, Investigation, Writing – original draft. **Yingfei Sun:** Supervision, Writing. **Mengjie Ren:** Investigation, Writing – original draft of Section 4.3–4.5, Software. **Shiguang Guo:** Investigation, Writing – original draft of Section 4.1–4.2, Software. **Ruotong Pan:** Investigation, Writing – original draft of Section 3.2.3. **Hongyu Lin:** Conceptualization, Investigation, Writing – review & editing. **Le Sun:** Supervision. **Xianpei Han:** Conceptualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

We sincerely thank all reviewers for their insightful comments and valuable suggestions. This research work is supported by the National Natural Science Foundation of China under Grants no. 62122077, 62106251, 62306303. Xianpei Han is sponsored by CCF-BaiChuan-Ebtech Foundation Model Fund.