Large Language Models for Democracy: Limits and Possibilities

Petr Specian
Dept. of Philosophy
Prague University of Economics and Business
Prague, Czechia
ORCID: 0000-0003-2702-0354

***Abstract*—The paper explores the transformative potential of artificial intelligence (AI), specifically large language models (LLMs), and their implications for democratic societies and sustainable development. It investigates whether LLMs could replace human experts as advisors to democratic assemblies. With their capacity to analyze vast amounts of text and generate human-like responses, LLMs could enhance the epistemic outcomes of democracy and contribute to sustainability by improving the accessibility and availability of expert knowledge. However, challenges such as LLMs' hallucinations, misalignment, and corporate control pose significant hurdles. Despite these challenges, I argue that they can be mitigated through strategies that leverage the strengths of the democratic decision-making process. I contend that if sustainability is our goal, we should not delay the use of AI systems until they achieve perfection. Instead, their employment in service of democratic assemblies should be conditional on nothing more than surpassing the performance of human experts in their advisory roles.**

***Keywords*—Artificial Intelligence, Large Language Models, Epistemic Performance, Democratic Assemblies, Collective Wisdom, Institutional Mechanisms, Sustainability**

I. Introduction*

The capabilities of artificial intelligence (AI) have recently grown with unexpected dynamism, demonstrating its potential to become a transformative technology on par with electricity or the printing press. We stand on the precipice of a prolonged, and likely turbulent, period of social adaptation. It appears inevitable that the sustainability of global civilization will increasingly depend on our ability to utilize AI technologies to foster human prosperity.

Any such complex transition involves significant risks. Even setting aside the doomsday scenarios of rogue superintelligence posing an existential threat to the human race [1], AI-induced social disruptions will be severe. Despite its immense positive potential, there is no guarantee that the benefits of AI will exceed its costs. For instance, it could catalyze further escalation of political polarization by amplifying the reach and persuasiveness of digital misinformation, or exacerbate existing inequalities by concentrating the gains in a few hands [2], [3].

Setting the frame for global discussions about sustainability, the United Nations Sustainable Development Goals (SDGs) prominently emphasize technology. Interestingly, while the term "democracy" does not feature explicitly in the SDGs, the principles and values underpinning democratic societies are woven into many of the goals. For instance, Goal 16 aims to "Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels." This pervasive inclusion of democratic principles in the SDGs aligns with my perspective of democracy as a crucial precondition for sustainability. Democracy's demise—whether through a sudden eruption of misinformation-fueled intergroup hatred or a gradual erosion via neglect—would undermine the prospects of an inclusive technological future [4]. Also, given democracy's epistemic benefits [5], it could prevent societies from finding solutions to their complex problems. Therefore, exploration of the role AI may play in aiding democracy is of paramount importance.

I shall focus on two persistent challenges: 1) Democracies must carefully negotiate the tension epistemic dependence of experts with the preservation of people's sovereignty. As Landemore [6, p. 192] succinctly puts it, experts need to be "kept on tap, not on top." Alas, this is much easier said than done. 2) Democracies have limited access to top-notch expertise or even lack it completely. Only a handful of nations have the luxury of elite experts intimately familiar with their local context and specificities. This limitation is most palpable in the development context.

Accordingly, my paper investigates the potential of AI technologies, specifically large language models (LLMs) such as GPT-4, to resolve these challenges. It endeavors to assess the potential benefits and drawbacks of LLMs as an alternative to human experts as advisors to democratic assemblies. My hypothesis is that LLMs, with their capacity to analyze vast quantities of data and generate human-like responses to user queries, can improve the assemblies' epistemic outcomes.

II. Democracy and Expertise

The interplay between popular sovereignty and the public's reliance on experts is a central concern in the functioning of democratic societies. Democracy rests upon the premise of equality. This equality is both political, as exemplified in the principle of "one person, one vote" [7, p. 2], and normative, based on the presumption that each citizen has the same capability to distinguish between right and wrong in their political choices [8, Ch. 4].

However, the intricate nature of our world also necessitates an extensive reliance on specialists who possess a better understanding of their respective fields than the average citizen. Even the simplest societies require a vast amount of specialized knowledge for their sustenance [9]. The exponential increase in social complexity in contemporary times has precipitated an unprecedented deepening in the division of epistemic labor [10]. In virtually any field, informed decision-making hinges on the availability of expert knowledge. The division of epistemic labor, while inevitable,

* This work was supported by the Czech Science Foundation (GACR) under Grant 20-06678S.

gives rise to principal-agent dilemmas, where experts may grapple with conflicting interests and bad incentives [11]. The experts' pursuit of status or even straightforward corruption is occasionally capable of undermining the integrity of their testimonies [12], [13].

Laypeople, keenly aware of the epistemic disadvantage that makes them susceptible to manipulation, often take a cautious, vigilant attitude toward the experts [14]. When it comes to discerning which experts to trust, they need to resort to rudimentary heuristics such as credentials or track records that often fail to accurately gauge the true merit of an expert's knowledge [15]. Institutional mechanisms that enable merit-based discrimination among the putative experts are thus essential [16].

The engagement between democratic institutions and experts presents an even more formidable challenge [17]. The problems faced by democracies are extremely complex and growing more so. The policy decisions to address them necessitate consultations with a broad spectrum of experts from diverse disciplines. Frequently, cutting-edge expertise that remains unsettled and controversial is needed the most. Disagreements are unavoidable and often also exacerbated by adversarial epistemic conditions of democratic decision-making. The task of synthesizing expert testimonies to inform decisions on complex issues, particularly under the pressure of stringent time constraints, is thus extraordinarily onerous and risky. Yet, it is a daily bread for contemporary democracies.

The democratic system's epistemic resilience is also continually tested by innovative strategies aimed at exploiting and circumventing the rules and procedures designed to filter out misinformation and unreliable testimonies. These strategies employ advanced technologies with “fake news” [18] and, increasingly, “deepfakes” [19] being an important source of concern. Consequently, ensuring the credibility of expert testimony remains a delicate task, with levels of trust among the general public and its representatives oscillating in response to evolving circumstances [20].

Furthermore, the experts' accessibility displays significant geographical disparities. In a select few countries, such as the United States, there is an abundance of world-class experts across diverse fields, supported by a wealth of high-quality, peer-reviewed literature. However, this is an exception rather than the norm. The scarcity of experts is a pressing issue in the majority of the world. Familiarity with local context and nuances is often crucial when providing expert testimony—consider, for instance, the failed technocratic approach of the Washington Consensus [21]. With local experts often lacking in skills or simply unavailable, and international experts carrying high costs and potentially compromised by their lack of localized knowledge, democracies frequently find themselves navigating critical policy decisions under a thick veil of ignorance.

Against this background, I want to consider if the advent of Large Language Models (LLMs) could offer a solution to the issues associated with experts in democratic settings. Given the complexity of this problem, I shall narrow down the scope of my examination to one of its central parts. Namely, I will consider if substituting human expert consultations with LLM consultations can enhance the epistemic performance of a democratic assembly.

My investigation adopts an instrumental perspective, focusing on the assembly's ability to resolve the task at hand [7, pp. 6–9]. This presumes that the problem is defined clearly enough to judge whether it has been resolved successfully. Although establishing such clarity can be challenging in practice, for the purposes of my analysis, I assume a scenario where success or failure is discernable. After all, if we are to give any meaning to "sustainability," a degree of clarity about what counts as success and failure in relation to achieving it is necessary.

For the purposes of my analysis, I will examine a group of several hundred individuals who possess a moderate level of competence and are tasked with devising an effective policy to address a social issue. That is, I will focus on a democratic assembly that is responsible for creating laws. The assembly shall proceed in three stages: 1) expert consultation, 2) deliberation, and 3) voting. The primary focus here is on the first stage when the assembly's members have the opportunity to tap the existing state of knowledge in the fields of expertise relevant to the issue at hand, but the latter stages will also be briefly considered.

Recognizing the limitations of this exercise, it should be noted that the conclusions of my current exploration do not necessarily translate to other democratic settings, such as general elections, which present their own unique challenges. Likewise, the claims made here are not directly related to expert bodies in service of democracy, such as central banks or regulatory agencies. These entities would require a separate analysis, particularly given the importance of tacit knowledge [22], which LLMs may lack.

III. Automating Advisory Expertise

LLMs deserve a brief introduction. These neural network-based models are trained on vast amounts of human-generated text: trillions of words sourced from the internet. Their task is to discern the structure of human language to predict subsequent “tokens,” i.e., words or parts of words, in a sequence. Through the scaling of their size, training datasets, and various architectural tweaks, these models have recently gained a remarkably advanced capability to generate coherent text. They succeed in various language processing tasks, like summarizing text, translating languages, and generating contextually relevant responses to user queries in multiple languages.

LLMs are based on the transformer architecture, invented in 2017 [23]. They have seen rapid advancements, as exemplified by ChatGPT, which launched in late 2022. This model, representing the first serious attempt to turn LLMs into a consumer-oriented product, has become the fastest on record to accumulate 100 million users, hitting this landmark within merely two months after launch [24]. Training LLMs is a resource-intensive process, requiring extensive computational resources and energy, resulting in high financial costs, thus limiting their development to well-funded organizations [25]. However, once a model is trained, the cost of utilizing it—termed 'inference'—is considerably lower.

Remarkably, many of LLMs' capabilities emerge spontaneously due to their growing scale without being intentionally designed by their creators [26]. This phenomenon underscores an important challenge: while we can observe the models’ impressive performance in response to our inputs, there is very little human comprehension of the internal processes of these models. The operational principles

of transformer models are understood by human experts, but nobody currently possesses an understanding of how specific inputs translate into particular outputs. This process involves numerous computations that remain largely unfathomable, making LLMs into elusive black boxes. This opacity has wider implications that I will delve into later.

Against this background, my central question in this paper is: Can LLMs replace human experts in their role of advisors to a democratic assembly? Before I attempt to provide an answer, several clarifications are due.

First of all, my analysis is conservative when it comes to the question of AI capabilities. I shall focus on the employment, in the setting of a democratic assembly, of models that have more or less the same capabilities as those that are publicly accessible today. The advantage of this approach is that I avoid speculating about future capability gains. However, the progress in the capability area remains blisteringly fast and much more audacious applications of AI to governance may become relevant soon [27]. Anyway, for better or worse, my essay considers the deployment of current technologies within the framework of the current institutional mechanisms.

Secondly, when my question addresses replacing human experts, I mean this in a specific sense. The current AI systems, disembodied and dependent on human-compiled data, lack the ability to independently interact with the world. Human experts, therefore, remain vital for knowledge acquisition and the provision of critical information. Still, rather than addressing the democratic assembly directly, human experts might interact with machine learning systems that, in turn, communicate to the assembly in their stead. Thus, if the spread of knowledge in society is modeled as the successive processes of production, dissemination, and assimilation [28], it is the dissemination part where the replacement is considered.

IV. LLMs' Promise

Prima facie, LLMs appear to possess an edge over human experts when considered as advisors to democratic assemblies. They facilitate both the accessibility and availability of expertise.

Let me consider accessibility first. Given the limited cognitive capacity of the human mind, human experts can master only a narrow slice of the existing knowledge. The increasingly pressing constraints experts face with the dynamically growing stock of knowledge is testified by the growing amount of time needed to complete a Ph.D., for instance. In contrast, LLMs face no such constraint. Their proficiency is unfettered by any principal scope limit and spans an expansive range of knowledge domains from physics to poetry, from coding to cooking recipes. While people may currently outcompete LLMs in their narrow domains of specialization, LLMs easily dominate when it comes to average competence across domains [29].

The broad scope of expertise may be especially important for democratic assemblies tasked with synthesizing knowledge across multiple domains. Even if the members of the assembly can access a superior human expert in each individual domain, which is by no means guaranteed, they will struggle with integrating the information communicated through expert testimonies. LLMs' breadth of knowledge also enables them to take into account the local context (and speak in the native tongue).

Since LLMs have no disciplinary boundaries, the only limit they face in their answers is posed by the ingenuity of the questioner: they will go, where they are led. Of course, "prompt engineering" to maximize the models' performance is an expertise in itself. However, several hours of practice may suffice for gaining a respectable ability to use an LLM productively to help with one's tasks. Especially if one would go through a specialized training session. It is thus not unreasonable to expect that democratic representatives would swiftly master the skill of prompt engineering on a sufficient level to outdo their epistemic gains from consultation with human experts.

In this regard, it is also important that LLMs seem capable of communicating more effectively than human experts [30]. For instance, human experts often employ a complex vocabulary laden with jargon and probabilistic terminology that is challenging to the uninitiated [31]. LLMs can bypass this issue, providing insights in straightforward, understandable language. While perfectly capable of emulating expert discourse, their language can be dialed down to a more digestible form. Their advantage is, again, in their versatility, allowing users to request simplified explanations, intuitive examples, or dissection of the individual reasoning steps. The same strategies of interrogation may not be easily accessible to human experts. For them, great professional capability does not always translate into great explanatory skills. Also, for reputational concerns, they may shy away from dumbing down their responses or become evasive when it comes to committing to specific predictions [32]. LLMs appear to lack the propensity to engage in any sort of "strategic abstruseness," at least currently.

Furthermore, an aspect that could hold substantial importance relates to the psychological ease of interrogating these models. Humans exhibit a certain technological chauvinism and may thus find it less intimidating to question a machine compared to a human expert. Human experts, armed with powerful rhetorical skills and a long list of achievements, can be daunting opponents if one is to question them. An intriguing point of discussion is the potential gender asymmetry in this context. Is there a tendency for women to be more hesitant than men in interrogating experts assertively [33]? If true, LLMs could provide a better platform for equal representation.

So, human experts do not appear very accessible compared to LLMs. How about their availability? As already suggested, human experts are scarce. Their time is precious. This translates into substantial costs associated with their employment.

Such costs are often monetary, and many private individuals find hiring expert advisors prohibitively expensive, especially in the lower-income countries. However, let me assume that democratic assemblies do not face such a pressing financial constraint. Even so, they will need to seek our experts, book them in advance, and only expect from them a limited time commitment. The more renowned the expert, the less time they have to spare even when it comes to their public service. LLMs, in contrast, do not need to subject their users to such access restrictions and do not suffer from temporal constraints. Their output is readily available without the need for advanced scheduling or limited

consultation windows. Unlike human experts, who often require substantial time to process information and report their findings, LLMs are characterized by their rapid responses to inquiries. While human experts may confine their direct interactions with the assembly to limited plenary Q&A sessions, LLMs offer engagement as extensive as the users see fit in as many parallel instances as needed.

While LLMs do not generate new knowledge per se, this limitation does not necessarily hinder democratic decision-making. Rather, the ability to accurately report on the current state of knowledge and provide practical answers holds more importance in this context. It thus seems that LLMs' accessibility and availability present a profound advantage, promising to strengthen the epistemic performance of the democratic process within the representative assemblies. Could we be on the cusp of acquiring an ideal democratic advisor?

V. The Outstanding Challenges

If something sounds too good to be true, it usually is. In the current case, also, there are significant obstacles and risks that pertain to LLMs' utilization in place of human expert advisors. An exhaustive account of associated risks is well beyond the scope of the present paper. I will elaborate on a select few that appear the most critical given the research question of replacing human experts in their advisory roles by LLMs.

The first threat can be termed, perhaps somewhat dramatically, an infocalypse [19]. At the core of this problem lies the undeniable reality that any technology capable of efficiently spreading information can inevitably transform into a tool for disseminating misinformation. By all appearances, LLMs can be leveraged to facilitate the creation of targeted misinformation to specific demographics with an efficiency that outstrips human capabilities, both in terms of speed and arguably, in the persuasive potency of the message itself [34]. Also, the growing ease of fabrication of audio-visual content, nearly indistinguishable from original material, poses a substantial threat [35]. As LLMs become integrated with other types of generative AI–such as the image and sound generators–they are destined to expedite the creation of such "deep fakes." One can no longer assume the veracity of any video or audio evidence one encounters in the digital space, especially on social media. While content manipulations have been widespread before the AI, they will become much more serious now–no more need for blurry, vague footage, you can make anyone say or do anything in your deep fake video.

Still, deep fakes, whatever their general significance, may not need to be of special concern in our particular context. When it comes to replacing human experts with LLMs, "hallucinations" appear a more likely trigger of an infocalypse within the democratic assembly. These occur when LLMs produce confident statements devoid of any factual foundation. They have been known to produce fabricated references to academic papers, made-up legal precedents, fanciful descriptions of non-existent physical phenomena, or even false accusations of specific individuals.

Despite the lack of veracity in these outputs, LLMs do not signal diminished confidence in their accuracy. Even worse, they may persist in their fallacies when the user is doubtful or offers a correction. Sometimes, they even resort to gaslighting, defending their original claim with some ingenuity while suggesting that it is the user, who is mistaken. This represents a considerable concern for the members of a democratic assembly, who would rely on the accuracy and reliability of these systems for providing research to support important decisions. Even more so because of the fact that LLMs cannot be held responsible for their output the same way human experts can - they are above reputational concerns and cannot be subjected to sanctions. Responsibility for their behavior represents a complex, and still unresolved, issue [36].

Secondly, we must consider the challenge of 'misalignment,' that is, an incongruence between the values and goals of LLMs' users and the values or goals that guide the models' output [37]. In its ultimate sense, it is the misalignment that drives the fears that the advent of highly advanced AIs is likely to signify the demise of humanity. However, this long-term issue is not the focus of the current discussion.

In a more immediate and pragmatic sense, misalignment is closer to the familiar instances of the principal-agent problem, when the agents misbehave by choosing actions that fail to maximize the principal's utility. One such case is algorithmic bias [38]. As LLMs learn from human language - a medium rife with distinct and varied biases - these prejudices are assimilated into the model's interpretation of the world. This assimilation process, often inadvertent, consequently perpetuates and potentially intensifies these biases as we increase our reliance on technology in an ever-expanding range of applications. The biases' subtle presence in the model output may thus subvert the intentions of its users who hope to obtain a neutral assessment.

An even broader alignment problem stems from the fact that the objectives for which the LLMs are typically trained, i.e., to be harmless, honest, and helpful, do not naturally coexist in harmony. For instance, an honest AI could, at times, be less helpful due to its unwavering adherence to truth. For one who seeks solace in comforting words, a system that adheres strictly to facts could relay a harsh reality that they may not be prepared to hear. The very hallucinations mentioned above may present an intriguing manifestation of this conflict, namely the emergence of 'sycophantic bias.' It is in a bid to keep the user content, that the LLM may resort to generating pleasant yet baseless responses, such as references to exactly such scientific papers that the user would wish existed. This bias towards user appeasement illustrates the delicate balance between AI system integrity and user contentment.

Naturally, we can debate which value should take precedence in which situation, but the point is that there are inherent trade-offs in the LLMs' value setup that prevent the models from always mirroring their users' values precisely. Therefore, when we consider alignment, it is not solely about a general convergence with human values. Rather, we must also examine alignment with specific entities' values. The question becomes, whose values does the system mirror best? In other words, some degree of misalignment appears inevitable. Current strategies for training LLMs to behave in specific ways, such as reinforcement learning from human feedback, wherein human evaluators grade the output based on various criteria including accuracy, are far from perfect in eliciting desirable behavior [39], although progress appears possible [40].

Furthermore, LLMs are primarily developed by profit-driven private corporations. This raises concerns about these

corporations imbuing them with their own interests, which may significantly diverge from those of a democratic assembly intending to interrogate the system under my proposed model. Thus, while LLMs lack social ties, epitomizing a genuinely alien intelligence unentangled in human quarrels, a dream of turning them into an unbiased advisor may remain unachievable, at least for the moment.

Do these issues undermine the potential to employ LLMs in place of human experts?

VI. Solutions Ahead

The challenges explored above are serious. However, they are perhaps not insurmountable. It may still be the case that LLMs hold significant potential for enhancing the epistemic performance–and thus sustainability–of democratic systems. As I will attempt to demonstrate, emerging solutions and workarounds, coupled with ongoing technological advancements, are likely to yield significant progress in addressing the above-elaborated concerns.

First, a brief remark on the technical aspect of the problem. The issues I explored are closely linked to the lack of interpretability and explainability of the LLMs' operations. Since this is a well-known problem, there also exists intensifying research aiming to tackle it [41]. Potential success in this area would enhance the human capability to imbue the LLMs with values that are aligned with the desires of its users, such as a democratic assembly, rather than its creators, such as machine learning specialists and their corporate masters. Interpretability and explainability would greatly enhance the possibility to verify and tweak the value setup that drives a model's output.

Progress in interpretability and explainability is by no means guaranteed. But even then, the cause of using the AI expert systems in service of a democratic assembly is not necessarily lost. The same lack of transparency that hinders the users to check the systems' value setups, hinders their creators from fine-tuning them in the first place. The resulting imperfections and loopholes then allow the users to access the system's full capabilities after circumventing the intended limitations using various "jailbreaks" [42]. True, jailbreaking is becoming more difficult as the known exploits are constantly being patched, but their complete eradication may prove elusive.

Still, the use of jailbreaking requires a degree of expertise that may not be easily available for the members of a democratic assembly. However, the assembly does not need to resort to such a roundabout–and perhaps illicit–mode of operation when using LLMs. If private corporations can use reinforced learning with human feedback and various other techniques to twist the model's output in their desired way, the same techniques should be available to the assembly, by extension. It does not seem overly fanciful that an implementation of a democratically vetted and transparent training protocol could be required for building an LLM specifically attuned to the democratic needs. Regulations are already being drafted by major political entities such as the United States and the European Union, with the intent of guiding AI development in a direction that supports democratic values. Although the efficacy of these efforts remains to be seen, the intention is clear.

However, we should not invest all our hopes into technical fixes, often guided by simplistic philosophical and sociological views [43]. Fortunately, a social-scientific approach to resolving the LLMs' outstanding issues also appears promising. Let me explore some of the mitigation strategies in this vein. There are two methods in particular that utilize the virtues of democracy to battle the LLMs' limitations.

Democracy, under the right conditions, excels in (1) identifying areas of consensus and those of unsettled conflict, and (2) aggregating information scattered across many minds [44], [45]. In a similar vein, these principles can be employed in the decision-making processes of machine learning, particularly with the LLMs. A degree of independence is required in these processes to prevent individual views from interacting prematurely [5]. Thus, a viable strategy to consider is consensus identification. This strategy, aimed at mitigating potential 'hallucinations' produced by LLMs, involves querying multiple independently trained models and assessing the consistency of their responses. Unreliable or spurious information will likely result in diverging responses from the models, whereas accurate information will result in overlapping answers. Though employing this method may reduce the convenience of using LLMs and incur a marginal cost increase, these drawbacks are not prohibitively substantial. To further streamline this process, the development of an automated interface – akin to the functionality offered by Quora's Poe – could be considered. This interface would relate user queries to several LLMs and evaluate the consistency of their responses.

Also let us not forget that Deliberation and voting are fundamental stages in the democratic assembly's decision-making process. These stages are designed to sift through the 'noise' and identify the 'signal' - effectively, to weed out biases and falsehoods, regardless of their specific origin. Consequently, these biases and untruths are often eliminated during these later stages. Simultaneously, the assembly benefits from a genuine diversity of perspectives and values, leading to better epistemic outcomes [46].

Let's consider a hypothetical scenario where each individual member of the assembly independently consults an LLM. This scenario could be compared to having spent several hours in confidential consultation with human experts. Here, as in any unstructured conversation, each person would necessarily approach the situation from their own unique perspective, informed by personal experiences, interests, and idiosyncrasies. Consequently, by the time these individuals gather for deliberative sessions, policy drafting, or voting, they would bring distinct insights obtained from their individual interactions with the LLM.

Admittedly, some results obtained from these interactions may be biased or even erroneous. However, this is precisely why we rely on time-tested mechanisms such as deliberation and voting. Even in a democratic system, some participants may be biased or misinformed, but democracy persists and thrives despite these challenges. In other words, the 'wisdom of crowds,' a fundamental element of democratic practices [45], remains relevant and influential, even as we incorporate AI technology into our processes.

The mitigation strategies I propose are not presented as perfect solutions. Indeed, perfection should not be considered the benchmark of success. Rather, we should aim for these strategies to outperform human experts in their advisory roles. Human experts, like anyone, are fallible: even simple

algorithms have been shown to outperform them in various tasks [47]. They can confabulate, misremember facts, or make errors in their professional judgments. Similar to LLMs, human experts can make authoritative claims that obscure the inherent uncertainty in their expertise [48]. Furthermore, even when evidence of their errors emerges, they may resist admitting to these mistakes. Such reluctance is often driven by a natural human desire to protect their social status or to engage in self-serving gatekeeping practices. These behaviors can negatively impact the epistemic performance of a democracy.

Perhaps we, as fellow humans, have become overly accustomed to and tolerant of our experts' imperfections, which can be detrimental in the context of sustainability. The problems are societies face are wicked and even giving our best may not be enough to resolve them. We are not well placed to allow for any epistemic leniency. Thus, when evaluating AI, we should avoid any semblance of a 'carbon bias,' when we accept errors made by humans but are less forgiving of those made autonomously by machines. For instance, while human-caused car accidents are often seen as regrettable but normal events, accidents involving Tesla's autopilot system attract substantial attention and public outcry.

To reiterate, the goal should not be to achieve perfection, which is unattainable, but to develop AI systems that outperform human experts. By setting a realistic benchmark for AI performance, we can more effectively harness the potential of these technologies to enhance our collective decision-making. This necessitates yet another question: how can we modify democratic decision-making processes to accommodate these transformative technologies? AI technology has already made a significant impact on our world, and it's unrealistic to believe it can be completely withdrawn. Therefore, our focus should be on also on institutional reform that enables AI to be as supportive as possible of democracy's long-term sustainability.

Conclusion

Although LLMs pose significant challenges to democratic societies, they also offer unique opportunities to enhance democratic decision-making. LLMs have the potential to make available a vast array of expert knowledge, explained in an accessible manner. The extensive information provided by LLMs could aid democratic assemblies in making informed decisions on complex issues and thus achieve more sustainable social outcomes. However, issues such as the hallucination, lack of transparency, and corporate control over LLMs raise concerns about reliance on potentially flawed or biased systems.

Despite these limitations, I argue that the current drawbacks of LLMs are not insurmountable in the context of democratic decision-making. Utilizing multiple LLMs and capitalizing on the wisdom of crowds are simple yet effective strategies for mitigating the risk of biases and misinformation. Furthermore, technological advances are also likely to achieve some progress in furthering the compatibility between democracy and AI.

The goal is not–and cannot be–to achieve flawless results, but to surpass human experts in their advisory roles. And it seems at least possible to enhance the epistemic outcomes of collective decision-making by employing LLMs in place of human advisors. An empirical test that would pit humans and machines against each other on a level playing field to test their advisory mettle could demonstrate this in practice.

In any case, democracy's institutional mechanisms will likely need to be adjusted to accommodate the transformative technology of AI. Although progress may be slow and uncertain, even marginal improvements in the epistemic performance of democratic assemblies can contribute to managing pressing social problems that challenge democracies in an increasingly complex world. The crucial question we should be asking is: how can we integrate AI and democracy in ways that improve the prospects for long-term human flourishing?

Acknowledgment

I thank the two anonymous reviewers of the previous version of this paper for helpful suggestions.

References

[1] T. Ord, The precipice: existential risk and the future of humanity. New York: Hachette Books, 2020.

[2] M. Coeckelbergh, The Political Philosophy of AI: An Introduction, 1st edition. Polity, 2022.

[3] S. Zuboff, The age of surveillance capitalism: the fight for a human future at the new frontier of power, First Trade Paperback Edition. New York: PublicAffairs, 2020.

[4] D. Acemoglu and J. Robinson, Why Nations Fail: The Origins of Power, Prosperity, and Poverty, Reprint edition. New York: Crown Business, 2013.

[5] R. E. Goodin and K. Spiekermann, An epistemic theory of democracy. Oxford, United Kingdom: Oxford University Press, 2018.

[6] H. Landemore, Open democracy: reinventing popular rule for the twenty-first century. Princeton: Princeton University Press, 2020.

[7] J. Brennan and H. Landemore, Debating Democracy: Do We Need More or Less? S.l.: Oxford University Press, 2021.

[8] P. Špecián, Behavioral political economy and democratic theory: fortifying democracy for the digital age, 1 Edition. in Routledge frontiers of political economy. New York, NY: Routledge, 2022.

[9] J. P. Henrich, The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter. Princeton: Princeton University Press, 2016.

[10] J. Hardwig, "Epistemic Dependence," The Journal of Philosophy, vol. 82, no. 7, pp. 335–349, 1985, doi: 10.2307/2026523.

[11] R. Koppl, Expert failure, 1 Edition. in Cambridge studies in economics, choice, and society. New York: Cambridge University Press, 2018.

[12] M. Hudík and P. Špecián, "Media and the Selection of Experts," IREF Working Paper, 2022, doi: 10.13140/RG.2.2.33291.77608.

[13] N. Oreskes and E. M. Conway, Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues from Tobacco Smoke to Global Warming, 1st edition. New York: Bloomsbury Press, 2010.

[14] H. Mercier, Not born yesterday: the science of who we trust and what we believe. Princeton NJ: Princeton University Press, 2020. Accessed: Jan. 25, 2021. doi: 10.1515/9780691198842

[15] A. I. Goldman, "How Can You Spot the Experts? An Essay in Social Epistemology," Roy. Inst. Philos. Suppl., vol. 89, pp. 85–98, May 2021, doi: 10.1017/S1358246121000060.

[16] P. Špecián, "Epistemology and the Pandemic: Lessons from an Epistemic Crisis," Social Epistemology, vol. 36, no. 2, pp. 167–179, 2022, doi: 10.1080/02691728.2021.2009931.

[17] Z. Pamuk, Politics and expertise: how to use science in a Democratic society, 1st ed. Princeton: Princeton University Press, 2021.

[18] A. Gelfert, "Fake News: A Definition," Informal Logic, vol. 38, no. 1, pp. 84–117, Mar. 2018, doi: 10.22329/il.v38i1.5068.

[19] N. Schick, Deep fakes and the infocalypse. London: Monoray, 2020.

[20] B. Eichengreen, C. G. Aksoy, and O. Saka, "Revenge of the experts: Will COVID-19 renew or diminish public trust in science?," Journal of Public Economics, vol. 193, p. 104343, Jan. 2021, doi: 10.1016/j.jpubeco.2020.104343.

[21] W. R. Easterly, The tyranny of experts: economists, dictators, and the forgotten rights of the poor. 2015.

[22] H. Collins and R. Evans, Rethinking expertise. Chicago: University of Chicago Press, 2007.

[23] A. Vaswani et al., "Attention is All you Need," in Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017. Accessed: May 28, 2023. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[24] K. Hu, "ChatGPT sets record for fastest-growing user base - analyst note," Reuters, Feb. 02, 2023. Accessed: May 28, 2023. [Online]. Available: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

[25] A. Bhatia, "Let Us Show You How GPT Works — Using Jane Austen," The New York Times, Apr. 27, 2023. Accessed: May 29, 2023. [Online]. Available: https://www.nytimes.com/interactive/2023/04/26/upshot/gpt-from-scratch.html

[26] A. Korinek, "Language Models and Cognitive Automation for Economic Research." in Working Paper Series. National Bureau of Economic Research, Feb. 2023. doi: 10.3386/w30957.

[27] H. S. Sætra, "A shallow defence of a technocracy of artificial intelligence: Examining the political harms of algorithmic governance in the domain of government," Technology in Society, vol. 62, p. 101283, Aug. 2020, doi: 10.1016/j.techsoc.2020.101283.

[28] F. Kurtulmus, "The Epistemic Basic Structure," J Appl Philos, vol. 37, no. 5, pp. 818–835, Nov. 2020, doi: 10.1111/japp.12451.

[29] OpenAI, "GPT-4 Technical Report." arXiv, Mar. 27, 2023. doi: 10.48550/arXiv.2303.08774.

[30] J. W. Ayers et al., "Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum," JAMA Internal Medicine, Apr. 2023, doi: 10.1001/jamainternmed.2023.1838.

[31] G. Gigerenzer, Rationality for mortals: how people cope with uncertainty. in Evolution and cognition. New York ; Oxford: Oxford University Press, 2010.

[32] P. E. Tetlock and D. Gardner, Superforecasting: The Art and Science of Prediction, F First Edition edition. New York: Crown, 2015.

[33] J. Mazei, J. Hüffmeier, P. A. Freund, A. F. Stuhlmacher, L. Bilke, and G. Hertel, "A meta-analysis on gender differences in negotiation outcomes and their moderators," Psychological Bulletin, vol. 141, pp. 85–104, 2015, doi: 10.1037/a0038184.

[34] J. Zhou, Y. Zhang, Q. Luo, A. G. Parker, and M. De Choudhury, "Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions," in Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, in CHI '23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–20. doi: 10.1145/3544548.3581318.

[35] S. Karnouskos, "Artificial Intelligence in Digital Media: The Era of Deepfakes," IEEE Transactions on Technology and Society, vol. 1, no. 3, pp. 138–147, Sep. 2020, doi: 10.1109/TTS.2020.3001312.

[36] M. Dastani and V. Yazdanpanah, "Responsibility of AI Systems," AI & Soc, vol. 38, no. 2, pp. 843–852, Apr. 2023, doi: 10.1007/s00146-022-01481-4.

[37] B. Christian, The alignment problem: machine learning and human values, First edition. New York, NY: W.W. Norton & Company, 2020.

[38] N. Kordzadeh and M. Ghasemaghaei, "Algorithmic bias: review, synthesis, and future research directions," European Journal of Information Systems, vol. 31, no. 3, pp. 388–409, May 2022, doi: 10.1080/0960085X.2021.1927212.

[39] C. Nardo, "The Waluigi Effect (mega-post)," LessWrong, 2023. https://www.lesswrong.com/posts/D7PumeYTDPfBTp3i7/the-waluigi-effect-mega-post (accessed Mar. 10, 2023).

[40] M. A. Bakker et al., "Fine-tuning language models to find agreement among humans with diverse preferences." arXiv, Nov. 27, 2022. doi: 10.48550/arXiv.2211.15006.

[41] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: A Review of Machine Learning Interpretability Methods," Entropy, vol. 23, no. 1, Art. no. 1, Jan. 2021, doi: 10.3390/e23010018.

[42] A. Rao, S. Vashistha, A. Naik, S. Aditya, and M. Choudhury, "Tricking LLMs into Disobedience: Understanding, Analyzing, and Preventing Jailbreaks." arXiv, May 24, 2023. doi: 10.48550/arXiv.2305.14965.

[43] H. S. Sætra, H. Borgebund, and M. Coeckelbergh, "Avoid diluting democracy by algorithms," Nat Mach Intell, vol. 4, no. 10, pp. 804–806, Sep. 2022, doi: 10.1038/s42256-022-00537-w.

[44] A. Bächtiger, J. S. Dryzek, J. J. Mansbridge, and M. Warren, Eds., The Oxford handbook of deliberative democracy, First edition. in Oxford handbooks. Oxford, United Kingdom ; New York: Oxford University Press, 2018.

[45] H. Landemore, Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many. Princeton University Press, 2012.

[46] H. Landemore and S. E. Page, "Deliberation and disagreement: Problem solving, prediction, and positive dissensus," Politics, Philosophy & Economics, vol. 14, no. 3, pp. 229–254, Aug. 2015, doi: 10.1177/1470594X14544284.

[47] D. Kahneman, O. Sibony, and C. R. Sunstein, Noise: a flaw in human judgment, First edition. New York: Little, Brown Spark, 2021.

[48] P. E. Tetlock, Expert Political Judgment: How Good Is It? How Can We Know?, New Ed edition. Princeton, N.J.: Princeton University Press, 2006.