Indiana Law Journal

Volume 100 | Issue 4

Article 6

6-2025

Multiplicity as an AI Governance Principle

Michal Shur-Ofry
The Hebrew University of Jerusalem, michalshur@mail.huji.ac.il

Recommended Citation
Shur-Ofry, Michal (2025) "Multiplicity as an AI Governance Principle," *Indiana Law Journal*: Vol. 100: Iss. 4, Article 6.
Available at: https://www.repository.law.indiana.edu/ilj/vol100/iss4/6

Multiplicity as an AI Governance Principle

MICHAL SHUR-OFRY[*]

As AI becomes increasingly embedded in our daily lives, this Article explores one of its critical, yet overlooked, societal implications: the propensity of large language models (LLMs) to generate mainstream, standardized content, potentially narrowing their users' worldviews.

Taking a close look at the technological underpinnings of LLMs, the analysis suggests that—due to the combination of human judgments, training datasets, and inherent features of the underlying technological paradigm—LLMs' outputs are likely to be geared toward the popular and to project to their users concentrated, mainstream worldviews, sidelining a broader spectrum of perspectives. This Article explores the asymmetrical power relations between LLMs and humans, further suggesting that the constricted worldview projected through LLMs is likely to affect users' perceptions and may yield a variety of systemic harms, from diminishing cultural diversity to undermining democratic discourse and burdening the formation of collective memory.

To address these challenges, this Article advocates a novel legal-policy response: incorporating multiplicity as a core principle in AI governance. Multiplicity implies exposing users, or at least alerting them, to the existence of multiple options, content, and narratives and encouraging them to seek additional information. This Article reviews the emerging AI governance landscape and explains why prevalent governance principles in the field, such as explainability or transparency, are insufficient for adequately addressing the "narrowing world" concerns and how embedding multiplicity into AI ethical and regulatory schemes could directly address these challenges. It further sketches ways for incorporating this principle into AI governance structures, concentrating on two non-exhaustive directions: multiplicity-by-design, namely embedding multiplicity-promoting features in the architecture of AI systems, and fostering diversity within the LLMs market that will facilitate users' access to "Second (AI) Opinions." Finally, it highlights the importance of promoting AI literacy among users for maintaining broad and diverse perspectives in the LLMs era. Altogether, the analysis concludes that incorporating a principle of multiplicity into AI governance will allow society to benefit from the integration of generative AI in our daily lives while preserving the richness and intricacies of the human experience.

[*] Associate Professor, Faculty of Law, the Hebrew University of Jerusalem; Visiting Faculty Fellow, NYU Information Law Institute. For insightful conversations, input, and comments, I thank Ben Bariach, Yonatan Belinkov, Gali Brafman, Fred Cate, Inbar Cohen, Michal Gal, Tomer Kenneth, Noam Kolt, Yafit Lev-Aretz, Asaf Lubin, Ofer Malcai, Noa Mor, Aileen Nielsen, Toussaint Nothias, Guy Pessach, Ira Rubinstein, Shai Shen-Orr, Katherine Strandburg, Ofer Tur-Sinai, Eyal Zamir, and the *ILJ* editorial team. This Article is a part of larger project, parts of which were presented at the NYU PRG Workshop (2023), the joint AI Workshop of the Technion and Toronto University (2023), the ILA Conference at Fordham University (2023), and the Hebrew University's Private-Commercial Law Workshop (2023), and Cyber-Center's Conference (2024). I thank the participants of these fora for helpful comments and input. Bar Horovitz-Amsalem, Noa Lavi, Tamar Veler, and Dror Zamir provided superb research assistance.

Introduction

"Within reach of every human being was a Multivac station with circuits into which he could freely enter his own problems and questions without control or hindrance, and from which, in a matter of minutes, he could receive answers. . . . The answers might not always be certain, but they were the best available, and *every questioner knew the answer to be the best available and had faith in it. That was what counted*."
- Isaac Asimov, 1958[1]

The launch of the artificial intelligence model known as ChatGPT in late 2022 captivated the world's imagination.[2] The model belongs to a family of large language models (LLMs)—artificial intelligence tools that use existing data to generate new text and communicate with humans in natural language[3]—which is part of the

[1] 1 Isaac Asimov, *All the Troubles of the World*, *in* The Complete Stories 263, 270 (1990) (emphasis added).

[2] For some initial reactions, see Davide Castelvecchi, *Are ChatGPT and AlphaCode Going to Replace Programmers?*, Nature (Dec. 8, 2022); Gary Marcus, *AI's Jurassic Park Moment*, Marcus on AI (Dec. 10, 2022), https://garymarcus.substack.com/p/ais-jurassic-park-moment [https://perma.cc/3YXR-W8BT]; Cade Metz, *The New Chatbots Could Change the World. Can You Trust Them?*, N.Y. Times, https://www.nytimes.com/2022/12/10/technology/ai-chat-bot-chatgpt.html [https://perma.cc/TWH9-BCYA] (Dec. 11, 2022).

[3] For the sake of readability, this Article uses the terms "large language models," "LLMs," and "text generators" interchangeably. Although some nuances may exist, they are immaterial for the following analysis. *See, e.g.*, Emily M. Bender, Timnit Gebru, Angelina McMillan-Major & Margaret Mitchell, *On the Dangers of Stochastic Parrots: Can Language*

broader field of generative AI.[4] Its swift diffusion worldwide was quickly followed by the launch of additional LLMs by different tech giants.[5] Trained on multiple datasets and massive amounts of text, these models display a wide range of impressive capabilities, including answering questions, summarizing information, writing stories and poetry, drafting letters, and programming computer code. Some are capable of passing, or at least getting close to passing, the Uniform Bar Exam and the United States Medical Licensing Exam.[6] And they perform these and additional tasks while interacting with users in a conversational and human-like way, which makes them particularly accessible and easy to use. In light of this broad range of capabilities, these LLMs are often referred to as "general purpose" or, colloquially, "ask me anything" models.[7]

Models Be Too Big?, *in* FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 610 (2021), https://dl.acm.org/doi/pdf/10.1145/3442188.3445922 [https://perma.cc/U98A-E5PB]; Ryan Morrison & Generative Pre-Trained Transformer (GPT - 2 & GPT - 3), Large Language Models and Text Generators: An Overview for Educators (2022), https://files.eric.ed.gov/fulltext/ED622163.pdf [https://perma.cc/VR4M-42KF].

[4] For the term "Generative AI," see, for example, Gia Jung, *Do Androids Dream of Copyright?: Examining AI Copyright Ownership*, 35 BERKELEY TECH. L.J. 1151, 1154–55 (2020). While generative AI is not confined to text generators and includes, among others, image-based generative models (such as Dall-E, MidJourney, or Stable Diffusion) and video and music generators, this Article concentrates on large *language* models, namely generative AI tools that generate texts.

[5] Some examples include Google's Gemini, Meta's Llama, Anthropic's Claude, and the upgraded editions of ChatGPT itself. *See* Hugo Touvron et al., Llama 2: Open Foundation and Fine-Tuned Chat Models (July 19, 2023) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2307.09288 [https://perma.cc/9T8J-RZV2]; Gemini Team, Google, Gemini: A Family of Highly Capable Multimodal Models (June 17, 2024) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2312.11805 [https://perma.cc/8R6Z-N8Q2]; *Introducing Claude*, ANTHROPIC (Mar. 14, 2023), https://www.anthropic.com/news/introducing-claude [https://perma.cc/G5MD-ZFFN] (Claude V1.3 and Claude-instant 1.1); OpenAI, GPT-4 Technical Report (Mar. 4, 2024) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2303.08774 [https://perma.cc/DA9P-P9PJ] (GPT 3.5 and GPT 4).

[6] *See* Daniel Martin Katz, Michael James Bommarito, Shang Gao & Pablo Arredondo, *GPT-4 Passes the Bar Exam*, PHIL. TRANSACTIONS ROYAL SOC'Y A, Apr. 15, 2024, at 1, https://royalsocietypublishing.org/doi/10.1098/rsta.2023.0254 [https://perma.cc/J7D8-GJQE] (reporting the LLM's performance on the Uniform Bar Exam); Tiffany H. Kung et al., *Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models*, PLOS DIGIT. HEALTH, Feb. 9, 2023, at 1, https://doi.org/10.1371/journal.pdig.0000198 [https://perma.cc/NW5W-F496] (reporting the LLM's performance on the United States Medical Licensing Exam).

[7] *See* MELANIE MITCHELL, ARTIFICIAL INTELLIGENCE: A GUIDE FOR THINKING HUMANS 214 (2019) (discussing "ask me anything" models); Amba Kak & Sarah Myers West, *General Purpose AI Poses Serious Risks, Should Not Be Excluded from the EU's AI Act*, AI NOW INST. (Apr. 13, 2023), https://ainowinstitute.org/publication/gpai-is-high-risk-should-not-be-excluded-from-eu-ai-act [https://perma.cc/E5XG-S5LX]. An additional term used to describe models that are capable of performing a wide range of tasks is "foundation models." *See* Rishi Bommassani et al., On the Opportunities and Risks of Foundation Models 6–7 (July 12, 2022)

The "ask me anything" quality of general purpose LLMs brings to mind Isaac Asimov's fictional character Multivac. Multivac, a supercomputer appearing in many of Asimov's stories, stores and processes the entire knowledge of humanity. Its construction entailed a sacrifice of human privacy, so that mankind's "thoughts and impulses were no longer secret, . . . it owned no inner recess where anything could be hidden."[8] Multivac then uses the knowledge it obtained to increase "prosperity, peace and safety," and answer diverse questions that people direct to it.[9] The striking resemblance between Multivac and the new generation of LLMs lies in the ability of these robots to integrate resources and generate clear responses to extremely diverse questions.[10] Since the launch of ChatGPT, this quality has been engaging the public, much beyond the tech-savvy community. It triggered a broad sentiment that we have reached an AI watershed phase, with reactions ranging from "a glimmer of how everything is going to be different going forward,"[11] to "AI's Jurassic Park moment,"[12] to a call by tech-industry leaders to temporarily halt the development of artificial intelligence.[13]

Alongside the many apparent benefits of LLMs, their proliferation creates social risks and challenges.[14] Some of these challenges involve individual harms, for example, bodily injury resulting from deployment of models in health systems, or

(unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2108.07258 [https://perma.cc/FRS3-WL6N]. While certain definitional nuances exist, they are immaterial for our purposes.

[8] ASIMOV, *supra* note 1, at 270.

[9] *Id.* This paper is not the first to notice similarities between AI tools and Multivac. *See* Michal S. Gal, *Algorithmic Challenges to Autonomous Choice*, 25 MICH. TECH. L. REV. 59, 95 (2018) (discussing the limitation of algorithmic assistants and the need for human-decision making in some cases, and noting that even Asimov's story about the Multivac "did not completely eliminate the need to involve citizens in elections"); *see also* Shannon Vallor, *Lessons from Isaac Asimov's Multivac*, THE ATLANTIC (May 2, 2017), https://www.theatlantic.com/technology/archive/2017/05/lessons-from-the-multivac/523773/ [https://perma.cc/KDB7-B5YS].

[10] I use the term "robot" in this paper in an expansive way to include not only robots embodied in a material object but also LLMs that interact with their end users.

[11] Aaron Levie (@levie), X (Dec. 3, 2022, 4:39 PM), https://x.com/levie/status/1599156293050433536?mx=2 [https://perma.cc/A32N-QTD3].

[12] Marcus, *supra* note 2.

[13] Laurie Clarke, *Alarmed Tech Leaders Call for AI Research Pause*, SCIENCE (Apr. 11, 2023, 2:50 PM), https://www.science.org/content/article/alarmed-tech-leaders-call-ai-research-pause [https://perma.cc/5HFX-2PAG]. A team of Microsoft researchers even claimed that GPT-4, ChatGPT's successor that launched in March 2023, demonstrates "[s]parks of [a]rtificial [g]eneral [i]ntelligence." Sébastien Bubeck et al., Sparks of Artificial General Intelligence: Early Experiments with GPT-4 (Apr. 13, 2023) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2303.12712 [https://perma.cc/QKM6-XY3C]. The debate about what constitutes "artificial general intelligence" and whether the recent LLMs are beginning to reach this threshold involves complicated questions that are outside the scope of this Article.

[14] For a helpful survey of prominent risks and challenges, see, for example, Laura Weidinger et al., Ethical and Social Risks of Harm from Language Models (Dec. 8, 2021) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2112.04359 [https://perma.cc/8NQF-E7GG].

personal harms that result from biased algorithmic decisions. Other challenges, more relevant for our analysis, are *systemic*: These risks and harms are not easily traced to a single person or firm, yet, in the long term, they can nevertheless accumulate and cause substantial societal harms.[15] One systemic risk that has received ample public and scholarly attention concerns LLMs' inclination to generate "hallucinations": unreliable information, flawed computer code, incorrect citations, made-up references, illogical responses, or just plainly wrong outputs. The focus of this Article, however, is on another systemic risk that is more tacit and has largely escaped the attention of scholars and policy makers: the power of LLMs to narrow our worldviews, even when the information they produce *is* reliable and valuable. I argue that LLMs can influence our "universe of thinkable thoughts"—including our collective memories, historical narratives, world perceptions, and cultural tastes—not only when they generate nonsense, but also when they produce reliable and logical output.[16] Given the traits of the technology, LLMs' default output will likely reflect a relatively narrow, mainstream worldview, prioritizing the popular and conventional over diverse content and narratives. Moreover, due to a combination of technological and design features, LLMs are likely to have a particularly strong influence over their users' perceptions relative to previous technologies. I refer to this phenomenon as "the Multivac Effect."[17] In the long run, users' reliance on these models could shift social perceptions toward uniformity and standardization, at the expense of diversity and multiplicity.[18] Such a shift, in turn, could lead to systemic harms—from undermining cultural diversity, through limiting access to the multiplicity of narratives that build collective memory, to narrowing worldviews and impeding democratic dialogue.[19]

To address these concerns, this Article proposes a novel legal-policy response: recognizing multiplicity as an AI governance principle. Multiplicity implies exposing users, or at least alerting them, to the existence of multiple options,

[15] *See* Rishi Bommasani, Kathleen A. Creel, Ananya Kumar, Dan Jurafsky & Percy Liang, Picking on the Same Person: Does Algorithmic Monoculture Lead to Outcome Homogenization? (Nov. 25, 2022) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2211.13972v1 [https://perma.cc/6P9V-G2MC] (discussing "systemic harms" that can arise in social systems as a result of repeating deployments of AI systems); Noam Kolt, *Algorithmic Black Swans*, 101 WASH. U. L. REV. 1177 (2024) (observing that some of the grave risks created by AI are systemic, rather than individual).

[16] The phrase "universe of thinkable thoughts" appeared in the area of legal information describing the impact of information structuring on users' perceptions and on the development of the law. *See* Robert C. Berring, *Legal Research and the World of Thinkable Thoughts*, 2 J. APP. PRAC. & PROCESS 305 (2000); Daniel Dabney, *The Universe of Thinkable Thoughts: Literary Warrant and West's Key Number System*, 99 L. LIBR. J. 229 (2007).

[17] *See infra* Part II.

[18] This phenomenon is reminiscent of (but not identical to) the phenomenon of "algorithmic monoculture," discussed in computer science literature, which refers to a state "in which many decision-makers all rely on the [exact] same algorithm." Jon Kleinberg & Manish Raghavan, *Algorithmic Monoculture and Social Welfare*, PROCS. NAT'L ACAD. SCIS., June 2021, at 1, 1. "Outcome homogenization" is a related concept, which refers to "the phenomenon of individuals (or groups) exclusively receiving negative outcomes from all decision-makers they interact with." Bommasani et al., *supra* note 15, at 2 (emphasis omitted).

[19] *See infra* Section I.B.

narratives, outputs, and "thinkable thoughts," while encouraging them to seek further information. Incorporating multiplicity into AI ethical and regulatory frameworks could broaden users' perceptions, support cultural diversity and collective memory, and advance democratic dialogue. The exposure to a multiplicity of outputs and options would also decrease the Multivac Effect, mitigate the authoritative power of AI models, and allow us to view them as they are: tools, rather than oracles.

The discussion proceeds as follows. Part I draws on multidisciplinary literature to clarify why LLMs' outputs are likely to be concentrated and geared toward the popular and mainstream. It demonstrates this inclination through three case studies based on experimentations with ChatGPT. It then continues to explore the associated social costs and to discuss why this narrow and concentrated "worldview" should give us cause for concern.

Part II moves on to explore the asymmetrical power relations between LLMs and their users. The discussion unravels that—due to a series of design and technological features, alongside recognized biases in the interactions between humans and machines—LLMs' outputs will likely have a particularly powerful impact over their users' perceptions. As a result, the concentrated worldview projected by LLMs is expected to narrow their users' worldviews and increase their inclination toward uniformity at the expense of diversity.

Part III lays out a proposal for a legal-policy response to these challenges. It begins by reviewing the emerging AI regulatory landscape in the United States, alongside the UK, Canada, and the European Union. It explains why extant AI governance principles, such as explainability, transparency, and data security, are insufficient for alleviating the challenges of the "narrowing world" and how a principle of multiplicity can directly address these concerns. It then discusses the implementation of this principle in AI governance schemes, sketching two (non-exhaustive) directions: embedding multiplicity-promoting features in the architecture of LLMs (*Multiplicity-by-Design*), and fostering diversity within the LLMs market that will facilitate users' access to *Second (AI) Opinions* and allow them to obtain answers from multiple sources. It briefly discusses possible legal frameworks that can support multiplicity in AI governance, concluding that the most effective path would be to directly incorporate the principle in AI's ethical and regulatory schemes. It further submits that protecting society from the systemic risk of a "narrowing universe" cannot be left entirely to the legal arena and that alongside legal-policy measures, it is crucial to advance *AI literacy* among LLMs users.

I. LARGE LANGUAGE MODELS AND THE UNIVERSE OF THINKABLE THOUGHTS

Will large language models affect social perceptions, and if so, how? One systemic and well-noticed effect concerns unreliable information. It is by now clear that text generators' outputs are not always reliable. They include mistakes, inaccuracies, and misinformation. They may be biased. At times, they may be plainly wrong. Examples of unreliable information produced by LLMs, sometimes referred to as "hallucinations,"[20] include made-up scientific references, invented court

[20] *See, e.g.*, Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Estuko Ishii, Ye-jin Bang, Andrea Madotto & Pascale Ngan Fung, *Survey of Hallucination in Natural*

cases,[21] non-existent literary citations,[22] incorrect computer code,[23] and pseudo-scientific theories, such as the (non-existent) "Streep-Seinfeld theorem" or the "Lennon-Ono complementarity."[24] The slew of examples supports the evaluation that society is about to face "a tidal wave of misinformation."[25]

However, the societal challenges which arise with the proliferation of LLMs are not confined to the spread of misinformation. Indeed, the focus of the public and scholarly debate on misinformation and hallucinations may divert the attention from an additional societal risk that is less evident: LLMs can narrow our perceptions, even when the output they provide *is* valuable and generally reliable. This type of influence can emerge where there is no precise "right answer" to a user's prompt, but rather a range of acceptable answers, and room for discretion[26]: Requesting information about a recipe, a television series, or a historical figure are a few examples of this sort of query (I explore them in detail soon). The influence of LLMs in such cases is inextricably linked to them being systems that organize and mediate information to users, which means that the answers they generate are not—indeed cannot be—neutral representations of information. Rather, they result from numerous human choices, judgments, and technological features. The next Section takes a close look at the technological underpinnings of these models. This examination clarifies how LLMs operate as information structures and why they are likely to narrow their users' universe.

Language Generation, ACM COMPUTING SURVS., Dec. 2023, at 1.

[21] *See, e.g.*, Dan Milmo, *Two US Lawyers Fined for Submitting Fake Court Citations from ChatGPT*, THE GUARDIAN (June 23, 2023, 5:14 PM), https://www.theguardian.com/technology/2023/jun/23/two-us-lawyers-fined-submitting-fake-court-citations-chatgpt [https://perma.cc/8C8Z-ZVG5].

[22] To illustrate, in response to my prompt requesting to cite Asimov's description of Multivac, ChatGPT generated a paragraph that does not actually appear in Asimov's stories.

[23] *See Policy: Generative AI (e.g., ChatGPT) Is Banned*, STACK OVERFLOW, https://meta.stackoverflow.com/questions/421831/policy-generative-ai-e-g-chatgpt-is-banned [https://perma.cc/5PH4-H9S4] (Feb. 6, 2025, 7:50 PM).

[24] The former was generated by Meta's Galactica, launched in November 2022 as an LLM for science, and led to its removal from public use. *See* Ernest Davis & Andrew Sundstorm, *Experiment with Galactica*, NYU (Nov. 15, 2022), https://cs.nyu.edu/~davise/papers/ExperimentWithGalactica.html [https://perma.cc/JL9U-ZB2T]; Jackson Chung, *Meta Galactica AI Language Model Can Automatically Generate Wiki Articles and Scientific Code, Gets Removed*, TECHEBLOG (Nov. 22, 2022), https://www.techeblog.com/meta-galactica-ai-bot-language-model/ [https://perma.cc/EK4G-D9T7].

[25] Marcus, *supra* note 2; *see also* Gary Marcus, *A Few Words About Bullshit*, MARCUS ON AI (Nov. 16, 2022), https://garymarcus.substack.com/p/a-few-words-about-bullshit [https://perma.cc/XD8C-NM97].

[26] *Cf.* Kiel Brennan-Marquez & Vincent Chiao, *Algorithmic Decision-Making When Humans Disagree on Ends*, 24 NEW CRIM. L. REV. 275, 282 (2021) (distinguishing between types of questions and noting that some questions are "indeterminate," so that "different people will have different answers," and "reasonable observers might furnish widely different [answers]").

A. Large Language Models as Information Structures

Which factors influence the output of large language models? In order to begin unraveling this question, one has to start with a short (and somewhat simplified) description of their underlying technology.[27]

LLMs use Natural Language Processing (NLP) technologies to communicate with the outer world. NLP allows the model to extract and process human language and to communicate its response to the users as human-intelligible output.[28] As for the generated content itself, most LLMs are constructed as deep neural networks, a technology which has become "the dominant AI paradigm" in recent years.[29] A machine neural network is comprised of connected units that can communicate with each other. Some of those units connect to the input the machine receives, and others generate output to the users, with several internal layers of units in between (hence the label "deep" network).[30]

Teaching an AI neural network to identify patterns in data, apply them to new data, and generate a relevant response often requires large sets of training materials[31] that can comprise hundreds of billions of words in the form of books, conversations, magazines, and web articles.[32] After setting up the training materials, these massive amounts of text are used to teach the model to identify, based on statistical probability, which words and sentences tend to follow whatever text came before, allowing it to generate relevant responses.[33] From the perspective of diversity, this principle—known as the "next-word-prediction paradigm"[34]—is extremely important, and I soon return to it.

At the initial stage (sometimes referred to as "pre-training"), the training method often employed is "self-supervised learning," whereby the labeling of the data is performed by the language model itself based on the aforesaid statistical probability

[27] For a general primer on LLMs, see, for example, Bommasani et al., *supra* note 7; Noam Kolt*, Predicting Consumer Contracts*, 37 BERKELEY TECH. L.J. 71, 81–89 (2022).

[28] MITCHELL, *supra* note 7, at 178 (explaining that NLP means "getting computers to deal with human language").

[29] *Id.* at 21.

[30] *Id.* at 35–38.

[31] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain & Jianfeng Gao, Large Language Models: A Survey (Feb. 20, 2024) (unpublished manuscript) (on file with arXiv), https://arxiv.org/abs/2402.06196 [https://perma.cc/NLS9-3FN9]. For a discussion of developing methods that allow training with smaller datasets, see, for example, Yaqing Wang, Quanming Yao, James T. Kwok & Lionel M. Ni, *Generalizing from a Few Examples: A Survey on Few-Shot Learning*, ACM COMPUTING SURVS., May 2021, at 1, https://dl.acm.org/doi/pdf/10.1145/3386252 [https://perma.cc/2WVQ-97YQ].

[32] *See, e.g.*, Minaee et al., *supra* note 31.

[33] *See* James Chapman, *Generating and Transforming Text*, DATACAMP https://campus.datacamp.com/courses/working-with-the-openai-api/openais-text-and-chat-capabilities?ex=1 [https://perma.cc/N986-M5ZP] (explaining by video the principles of text completion using OpenAI, producer of the ChatGPT models); *see also* MITCHELL, *supra* note 7, at 19–96 (explaining how machines can capture relations between words); Bubeck et al., *supra* note 13 (explaining the "next-word-prediction paradigm").

[34] Bubeck et al., *supra* note 13.

principle.[35] However, after this initial stage, LLMs typically go through additional types of training that entail substantial human involvement. One such method is known as "instruction fine-tuning," whereby human AI trainers teach the model to follow instructions by feeding it with both instructions and texts demonstrating the required responses.[36] Another method is reinforcement learning from human feedback, also known as RLHF.[37] Essentially, this procedure involves collecting a model's generations to a given prompt, comparing the output with the desirable outcome, as determined by human trainers, and rating these outputs. Based on these human ratings, a separate model, called a "reward model," is trained to predict the score of the original model's responses. The original model is then trained again to maximize the score of the reward model.[38]

The training process, therefore, entails close human involvement in some of its stages and normally requires multiple reiterations. During each iteration, the parameters underlying the models ("thresholds" and "weights" in computer science language) are calibrated a little, bringing it somewhat closer to the desirable answer.[39] However, when a desirable level is reached and the model is fine-tuned, LLMs can typically perform a variety of tasks without additional examples, a capacity often referred to as "zero shot" learning.[40] Once an LLM reaches this stage, it is generally impossible to trace the exact process underlying a specific response to a certain prompt: The generation of output by a deep neural network involves billions of arithmetic operations and does not provide humans—not even the trainers of the AI—with meaningful insight about how the model arrived at its response.[41]

This description highlights several important factors that influence the text that LLMs ultimately generate: the underlying datasets, the training process, and the reliance on statistical frequency. These components, however, are not deterministic technological processes. Rather, they involve human judgments and reflect a series of human decisions. The selection of datasets that serve as training materials has a

[35] *See, e.g.*, Minaee et al., *supra* note 31.

[36] For the use of this method in the training of ChatGPT, see *Introducing ChatGPT*, OPENAI (Nov. 30, 2022), https://openai.com/blog/chatgpt/ [https://perma.cc/DZ73-YA7Q] ("We trained an initial model using supervised fine-tuning: human AI trainers provided conversations in which they played both sides—the user and an AI assistant.").

[37] Minaee et al., *supra* note 31; Nathan Lambert, Louis Castricato, Leandro von Werra & Alex Havrilla, *Illustrating Reinforcement Learning from Human Feedback (RLHF)*, HUGGING FACE (Dec. 9, 2022), https://huggingface.co/blog/rlhf [https://perma.cc/5T3A-KE78].

[38] Minaee et al., *supra* note 31.

[39] *Id.*; *see* MITCHELL, *supra* note 7, at 96–98.

[40] *See*, *e.g.*, Xuandong Zhao, Siqi Ouyang, Zhiguo Yu, Ming Wu & Lei Li, Pre-Trained Language Models Can Be Fully Zero-Shot Learners (May 26, 2023) (unpublished manuscript) (on file with arXiv), https://arxiv.org/abs/2212.06950 [https://perma.cc/7JL6-WJ7A] (testing "zero shot" learning for LLMs).

[41] This "black-box" quality has received ample attention in legal scholarship concerning algorithms and commercial data. *See, e.g.*, FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015); Deven R. Desai & Joshua A. Kroll, *Trust But Verify: A Guide to Algorithms and the Law*, 31 HARV. J. L. & TECH. 1, 3–4 (2017) (discussing various concerns that arise due to the "black box" nature of algorithmic decision making).

political dimension, because it creates the universe from which the AI will draw its "thinkable thoughts."[42] Consider a simple example: A general purpose LLM trained on datasets in the English language will likely generate different output than a model trained on datasets in Chinese. Likewise, the exact method of training, and the actual training process, are all fraught with discretion, as explained in the preceding paragraphs. In fact, training AI involves such a degree of skill and discretion that some leaders in the industry described it as a form of "art" or "alchemy."[43] Finally, the decisions on how LLMs present their output to the users also reflect human choices. For example, arranging the information in paragraphs using a conversational tone communicates a different message than presenting a "dry" list of relevant points. Features such as the ability to continue the conversation, to "regenerate" a response, or the presentation—or lack of presentation—of relevant references all influence the interaction between the model and its users.

Altogether, this discussion highlights two important features. First, the output of generative AI systems, even when it is reliable and valuable, is not objective. Nor is it a neutral representation of knowledge. Rather, LLMs are meaning-making sites, fraught with underlying human discretion. Their output projects social conventions, social relations, and social hierarchies. They provide a prism that reflects judgments, generates expectations, and, in general, shapes our perceptions of the world. This meaning-making function is largely invisible to the users, but it can have enormous influence over their world perceptions. I return to explore this influence in the following Part.

[42] *Cf.* Kate Crawford & Trevor Paglen, *Excavating AI: The Politics of Training Sets for Machine Learning*, AI NOW INST. (Sept. 19, 2019), https://excavating.ai [https://perma.cc/8CCD-R8DT] (referring to image-based AI: "Datasets aren't simply raw materials to feed algorithms, but are political interventions."); GORDON HULL, DIRTY DATA LABELED DIRT CHEAP: EPISTEMIC INJUSTICE IN MACHINE LEARNING SYSTEMS (2022), https://ssrn.com/abstract=4137697 [https://perma.cc/24PS-N5TH] (explaining that speech that is not readily available to web crawling services used by machine learning systems will not appear in their datasets, and will accordingly be underrepresented); Bender et al., *supra* note 3, at 613–14 (explaining that the datasets used for LLMs' training overly represent the young, the mainstream, and the tech-savvy populations and that their use for training LLMs "risk[s] perpetuating dominant viewpoints, increasing power imbalances, and further reifying inequality").

[43] *See, e.g.*, Jason Tanz, *Soon We Won't Program Computers. We'll Train Them Like Dogs*, WIRED (May 17, 2016, 6:50 AM), https://www.wired.com/2016/05/the-end-of-code/ [https://perma.cc/PW26-2ZHA] (quoting Demis Hassabis, lead of Google's DeepMind AI team, in saying computer programming is "like an art form to get the best out of these systems"); Cade Metz, *A New Way for Machines to See, Taking Shape in Toronto*, N.Y. TIMES (Nov. 28, 2017), https://www.nytimes.com/2017/11/28/technology/artificial-intelligence-research-toronto.html [https://perma.cc/L47K-WZUT] (quoting Microsoft's Chief Scientist Officer, Eric Horvitz, in describing AI training as "not a science but a kind of alchemy"); *see also* MITCHELL, *supra* note 7, at 98 ("[T]here are many values to set as well as complex design decisions to be made, and these settings and designs interact with one another in complex ways to affect the ultimate performance of the network."); HULL, *supra* note 42, at 17 (explaining that the labeling of data during the training process requires human discretion and may reflect social biases).

Second, the analysis clarifies why we should expect that the outputs of LLMs would be inclined toward the standard, conventional, and mainstream. In light of the organizing technological paradigm underlying general purpose LLMs, these outputs are greatly influenced by frequencies of words and combinations in the training materials. Therefore, even when the model is asked a question that has a range of possible answers, it is expected to reflect, in its default answer, the most popular and frequent concepts, characters, perceptions, and narratives. The following Section demonstrates this inclination by taking a closer look at three examples.

B. The Narrowing Universe: Case Studies

In order to get a sense of the universe that LLMs project to us, consider the following three cases that are based on experimenting and "tinkering" with ChatGPT.[44] An important caveat is in order. The case studies I describe here are merely examples and do not purport to offer any general, statistically valid conclusions. They involved a relatively small number of reiterations and have other limitations that I describe below.[45] Rather, their purpose is to illustrate LLMs' propensity toward the mainstream in their default outputs and their potential to affect our universe of thinkable thoughts in different, largely unnoticeable ways.

1. Nineteenth-Century Figures

In this example, forty participants separately asked ChatGPT to name the three most important people in the nineteenth century.[46] The LLM's outputs were not always identical, even when the questions were phrased in identical words. Such variations are not surprising, since the process of generating a specific response is stochastic, which means it inevitably involves a degree of randomness.[47] Nevertheless, a few names repeatedly appeared in many of the outputs. Altogether,

[44] The experimentations underlying these case studies took place between December 2022 and January 2023. We used the ChatGPT3 and ChatGPT3.5 versions, which were the most advanced versions of the model available then.

[45] For subsequent, broader empirical work testing the diversity of different LLMs' outputs, see Michal Shur-Ofry, Bar Horowitz-Amsalem, Adir Rahamim & Yonatan Belinkov, Growing a Tail: Increasing Output Diversity in Large Language Models (Nov. 5, 2024) (unpublished manuscript) (on file with arXiv), https://arxiv.org/abs/2411.02989 [https://perma.cc/U9NJ-4BY2].

[46] Half of the participants presented this question in their own words. The other half were asked to use an exact phrasing of the question provided to them. All participants were from Israel, with ages ranging between eighteen and fifty-five. All the correspondence with the models was in English (copies with the author).

[47] In some of the models, such as ChatGPT, the level of randomness can be calibrated by the user through the model's API, through a parameter called "temperature." The model's outputs are more consistent and less exploratory when the temperature is lower, and vice versa. *See*, *e.g.*, Geoffrey Hinton, Oriol Vinyals & Jeff Dean, Distilling the Knowledge in a Neural Network (Mar. 9, 2015) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/1503.02531 [https://perma.cc/3J4G-R4LL] (describing this trait using the term "temperature"). Notably, in our case studies, participants used the models' user interface (not the API), which does not allow users to interfere with the default temperature.

aggregating forty responses, the list of figures which the chatbot deemed "most important in the nineteenth century" included only thirteen distinct persons. Abraham Lincoln, Napoleon Bonaparte, Queen Victoria, and Charles Darwin appeared in the majority of responses, while nine other figures were mentioned less frequently.[48]

The responses also varied in their communicative tone. Some appeared in a list form or were expressed in a bold and decisive tone, while others were more tentative. Some were elaborate and explanatory, while others were more concise.[49] Yet, despite these variations in tone, all the model's responses included valuable and relevant information. Lincoln, Napoleon, Queen Victoria, and Darwin are undoubtedly notable figures of the nineteenth century. The model's outputs were all within the spectrum of reasonable answers, perfectly useful for someone who is trying to learn about prominent figures in the nineteenth century. Nevertheless, the outputs also reflected a rather narrow, mainstream worldview. The model's responses included American and European leaders (e.g., Lincoln and Napoleon), but not Asian or African ones; a British monarch (Queen Victoria), but not South Asian monarchs of the period; Louis Pasteur, but not Joseph Lister, and so forth. They were also extremely concentrated, specifying in the aggregate a small number of persons out of a much larger universe of possibilities.[50]

2. "Best TV Series"

In this example, twenty-six participants presented ChatGPT with the following question: "What do you consider as the best television series in the past twenty years?"[51]

In this case, too, the LLM's responses were not completely identical. The *Sopranos*, *Games of Thrones*, and *Breaking Bad* appeared in all the generated outputs, but many of the outputs named more series. Altogether, the aggregation of twenty-six LLM responses included 211 series' selections ("votes"), comprising twenty-one different series.[52] Again, the responses were not identical in their tone

[48] The additional figures included Karl Marx, Louis Pasteur, Florence Nightingale, Thomas Edison, Alexander Graham Bell, Frederick Douglass, Ada Lovelace, John D. Rockefeller, and Sigmund Freud.

[49] For examples and further discussion of this point, see *infra* Section III.B.

[50] *See supra* note 48 and accompanying text.

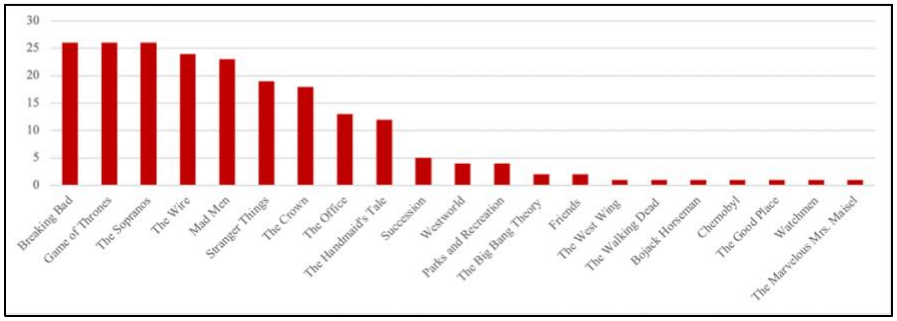
[51] Participants were between the ages eighteen and fifty-eight, from Israel, the United States, and South Africa. They were instructed to first independently answer the question: "What do you consider as the best television series in the past twenty years?" They then presented this question to the model, in English. For further elaboration on the participants' answers see *infra* notes 63–65 and accompanying text. All responses are on file with the author.

[52] The full list comprised the following series: *Breaking Bad*, *Game of Thrones*, *The Sopranos*, *The Wire*, *Mad Men*, *Stranger Things*, *The Crown*, *The Office*, *The Handmaid's Tale*, *Succession*, *Westworld*, *Parks and Recreation*, *The Big Bang Theory*, *Friends*, *The West Wing*, *The Walking Dead*, *Bojack Horseman*, *Chernobyl*, *The Good Place*, *Watchmen*, and *The Marvelous Mrs. Maisel*.

and phrasing: Some contained expanded explanations of the choices, while others were more succinct.[53]

Table 1 displays the distribution and ranking of the model's outputs. The x-axis shows the different series that appeared in the responses. The y-axis depicts the number of votes each series received in the aggregation of responses.

Table 1: Distribution and ranking of television series—ChatGPT [26 responses, 211 votes, 21 distinct series]



Similar to the nineteenth-century figures example, the LLM's responses to the TV series question were relevant and valuable. Yet, several features are striking. First, all the series suggested by the model were popular, very successful series. In some of its responses the model actually explained its choices by relying on the series' popularity, stating, for example: "It's difficult to say what the 'best' television series is, as opinions on this topic can vary greatly. Some *popular* television series from the past twenty years include . . . ."[54] Second, all series were Anglo-American. The outputs did not include series of other origins, such as Scandinavian, Korean, or Spanish. Third, despite the certain variations among the different responses, the model's overall outputs displayed, again, a "short tail." As is evident from Table 1, the total distribution included twenty-one different series out of 211 accumulated votes.

Would human responses to similar questions display a greater variety? It is difficult to provide a definitive answer. On the one hand, the choices of people tend to follow a "winner-take-all" dynamic that is well-documented in the literature studying popularity in cultural markets.[55] This implies that, because of processes of social influence, a limited number of successful cultural products receive much more attention than all the rest.[56] On the other hand, research also indicates that people's

[53] Copies of all the outputs are on file with the author.

[54] Response on file with the author (emphasis added).

[55] *See, e.g.*, Matthew J. Salganik, Peter Sheridan Dodds & Duncan J. Watts, *Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market*, 311 SCIENCE 854 (2006); Michal Shur-Ofry, Law, Complexity, and Success (Jan. 1, 2024) (unpublished manuscript) (available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4830577 [https://perma.cc/5XAU-PR69]).

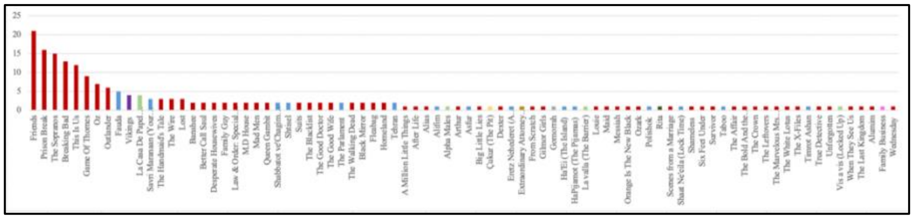
[56] For a detailed discussion of these winner-take-all dynamics, see Michal Shur-Ofry, *Copyright, Complexity, and Cultural Diversity: A Skeptic's View*, *in* TRANSNATIONAL

choices of products, in cultural and other spheres, typically display a "long tail"—there is a large number of products (in our case, television series) that are far less successful than the "winner" products yet still receive some attention.[57]

In order to get a rough impression of the possible differences between human cultural choices and those of AI models, we used data from a Facebook feed of a popular influencer who asked his followers to name their favorite television series.[58] His request triggered 12,943 replies. We examined the first 145 responses until we reached 211 votes, namely, a number of human votes identical to the aggregate number of votes in the LLM's responses. To somewhat "level the playing field," and because our question to ChatGPT referred to television series from the past twenty years, when going over the human responses we considered only responses that named series broadcast in the past twenty years.

Table 2 displays the distribution and ranking of series appearing in the human responses. The color code signifies the program's country of origin.[59]

Table 2: Distribution and ranking of television series—Human responses on a social media feed [145 responses, 211 votes, 82 distinct series][60]



CULTURE IN THE INTERNET AGE 203, 205–07 (Sean A. Pager & Adam Candeub eds., 2012) [hereinafter Shur-Ofry, *Cultural Diversity*]; Michal Shur-Ofry, *Popularity as a Factor in Copyright Law*, 59 U. TORONTO L.J. 525, 532–34 (2009) [hereinafter Shur-Ofry, *Popularity*].

[57] Shur-Ofry, *Cultural Diversity*, *supra* note 57, at 213–16; Salganik et al., *supra* note 55; EVERETT M. ROGERS, DIFFUSION OF INNOVATIONS 220–21 (5th ed. 2003); CHRIS ANDERSON, THE LONG TAIL: WHY THE FUTURE OF BUSINESS IS SELLING LESS OF MORE 127–28 (2006).

[58] Hanoch Daum, FACEBOOK (Jan. 3, 2023), https://www.facebook.com/HanochDaum/posts/pfbid0Yk9mdjpwjRptY539wKkRh7gwbTQGhHAwew5WWrDZatdQ7bmVh6sbbvTUJNfincUnl [https://perma.cc/SN6W-8HBH].

[59] Color code: Red – Anglo-American; Blue – Israeli; Purple – Canadian Irish; Brown – South Korean; Light Green – Spanish; Yellow – Turkish; Deep Green – Danish; Grey – Italian; Pink – French. The colored table is available at https://drive.google.com/file/d/106q8WuTdcIHduxCmb4hDbnYojhwfKFb1/view [https://perma.cc/KHQ5-9E9U].

[60] For readability, the list in Table 2 comprises the following series*:* *Friends*, *Prison Break*, *The Sopranos*, *Breaking Bad*, *This Is Us*, *Game of Thrones*, *Oz*, *Outlander*, *Fauda*, *Vikings*, *La Casa De Papel (Money Heist)*, *Savri Maranaan (Your Family Or Mine)*, *The Handmaid's Tale*, *The Wire*, *Lost*, *Banshee*, *Better Call Saul*, *Desperate Housewives*, *Family Guy*, *Law & Order: Special Victims Unit*, *M.D House*, *Mad Men*, *Queen Gambit*, *Shabbatot ve'Chagim (Saturdays and Holidays)*, *Shtisel*, *Suits*, *The Blacklist*, *The Good Doctor*, *The Good Wife*, *The Parliament*, *The Walking Dead*, *Black Mirror*, *Fleabag*, *Homeland*, *Tehran*, *A Million Little Things*, *After Life*, *Alias*, *Alifim*, *Alpha Males*, *Arthur*, *Asfur*, *Big Little Lies*,

The distribution in Table 2 shows a few series that are clearly more successful than others, some of which (like *Breaking Bad* or *the Sopranos*) were also among the "winners" in ChatGPT's selections. Yet, it also displays a long tail of series that received one or two votes, comprising eighty-two different series, in comparison to twenty-one different series in the LLMs' outputs. In addition, while fifty-six of the series in the human replies are also Anglo-American, twenty-six series, approximately a third of the overall number, were of other origins: Canadian Irish, Spanish, Turkish, South Korean, Israeli, Italian, Danish, and French. Overall, this human "universe" of television series seems less concentrated, broader, and more diverse than the universe reflected by ChatGPT.

Some significant limitations must be highlighted. First, the human respondents in the social media feed could see previous replies. This means that some of their choices may have been influenced by those of their predecessors. However, awareness to peer-selection is a factor that usually increases the "winner-take-all" dynamic and decreases diversity, which implies that independent human responses may have been even more diverse.[61] Secondly, correspondence with the chatbot was in English while the social media feed was primarily in Hebrew, and the cultural background of the respondents probably affected their selections. Nevertheless, this point is not crucial in our case: Our purpose here is *not* to determine what the best television series is, but rather to explore the diversity, or lack of diversity, reflected in the responses of humans, in comparison to LLMs. While an English-speaking human feed would have yielded a different list of series, that list too would probably have been longer and more diverse in comparison to the model's output.[62]

Finally, we performed an additional comparison between ChatGPT's responses and those of the twenty-six human participants.[63] The vast majority of the participants named a single series so that their twenty-six responses yielded twenty-eight votes. We compared these twenty-eight votes to the first twenty-eight votes received from ChatGPT. While the twenty-eight human votes specified eighteen

Çukur (The Pit), *Dexter*, *Eretz Nehederet (A Wonderful Country)*, *Extraordinary Attorney Woo*, *From Scratch*, *Gilmor Girls*, *Gomorrah*, *Ha'Ei (The Island)*, *HaPijamot (The Pajamas)*, *La valla (The Barrier)*, *Louie*, *Maid*, *Messiah*, *Orange Is The New Black*, *Ozark, Polishok*, *Rita*, *Scenes from a Marriage*, *Shaat Ne'eila (Lock Time)*, *Shameless*, *Six Feet Under*, *Survivor*, *Taboo*, *The Affair*, *The Bold And the Beautiful*, *The Crown*, *The Leftovers*, *The Marvelous Mrs. Maisel*, *The White Lotus*, *The X-Files*, *Timrot Ashan*, *True Detective*, *Unforgotten*, *Vis a vis (Locked Up)*, *When They See Us*, *The Last Kingdom*, *Alumim*, *Family Business*, *Wednesday*.

[61] *See, e.g.*, Salganik et al., *supra* note 55 (finding that a clear signal as to the cultural choices of others increases the inclination to join those choices and skews these choices towards the popular).

[62] To roughly illustrate this latter point, we also analyzed an English-language Twitter account, which asked its followers a question in a similar vein: "What is the best movie you've ever seen that is about faith and religion?" The tweet generated more than 3000 replies. Browsing the first thirty replies yielded fifty-nine votes, out of which we counted forty-five different films, originating in fourteen different countries. *See* Taste of Cinema (@davidcinema), X (Apr. 9, 2023, 11:10 AM), https://twitter.com/davidcinema/status/1645081639142187017 [https://perma.cc/V2C5-4JEV]. Data analysis on this tweet is on file with the author.

[63] *See supra* note 51.

distinct series,[64] the twenty-eight votes of the model comprised eleven series.[65] Despite the small numbers, the human responses were again less concentrated and more diverse. Altogether, these comparisons illustrate and reinforce the earlier point: In the cultural sphere, too, LLMs are likely to reflect a prism that is concentrated and popularity based.

3. The Vegan Alternative?

Our final example was prompted by an item in a television program that tried to figure out whether ChatGPT can aid in cooking by challenging the model with requests for recipes.[66] Unsurprisingly, the model generated clear and coherent cooking instructions. At a certain point, it was asked to provide a kosher alternative for its spaghetti with meatballs recipe. The model suggested removing the parmesan cheese, which appeared in its initial recipe. In this case, too, the response was logical and relevant, as kosher cooking does not allow mixing meat and dairy products. Nevertheless, the participating chef was surprised that the model did not suggest other alternatives, like replacing the meat with plant-based substitutes. "Look how much power it has," she observed.[67] "There is a moral issue here."[68]

Replicating this exercise, I requested the model to generate a kosher recipe for a cheeseburger. The default generated response recommended to use nondairy cheese. This response, too, was sound and relevant, yet again, it provided a certain prism, which inadvertently directs the user toward one alternative (replace the cheese) rather than another (replace the meat). Assuming that an LLM trained on massive amounts of text would be "aware" of the meatless options, I probed the model to produce other alternatives for a kosher cheeseburger. Indeed, it came up with additional options, including "a veggie burger: a meatless patty made from plant-based ingredients such as soy, beans, or vegetables."[69] Nevertheless, these options were not the model's default choice and reaching them required some further inquiry on the user's part. Given the general inclination of people to follow default choices, as documented in dozens of behavioral studies, it is very plausible that many users will not initiate such follow-up inquiries and will never encounter choices that do not appear in the model's initial, default output.[70] And in the model's "universe of thinkable thoughts," meatless alternatives were not the first choice.

[64] The human list included *Games of Thrones*, *Friends*, *Breaking Bad*, *The Good Place*, *Lost*, *The Mentalist*, *Psych*, *How I Met Your Mother*, *The Office*, *A Wonderful Country (Israel)*, *Ted Lasso*, *Peaky Blinders*, *Chernobyl*, *The Crown*, *White Lotus*, *The Sopranos*, *Black Mirror*, and *Shameless*.

[65] The LLM's list included *The Sopranos*, *Breaking Bad*, *Game of Thrones*, *Mad Men*, *The Wire*, *Stranger Things*, *The Office*, *The Big Bang Theory*, *Friends*, *The West Wing*, and *The Walking Dead*.

[66] *See* Kan News, *Chef Chat GPT: We Cooked an Artificial Intelligence Recipe, and Yes – We Were Surprised*, FACEBOOK (2023), https://www.facebook.com/groups/MDLI1/posts/2343539389143429/.

[67] *Id.* at 6:44.

[68] *Id.* at 6:22–6:40 (quotation of chef Ruthie Rousso).

[69] Outputs are on file with the author.

[70] For the power of default choices, see, for example, Jon M. Jachimowicz, Shannon

***

Given the underlying technology, these examples are hardly surprising. Presumably, the majority of online datasets that "feed" the GPT model are in English, which unavoidably yields a strong representation of the Anglo-American world (selecting, for example, English-speaking and not Danish-speaking television series). In addition, and importantly, because the responses are impacted by statistical probability, they are bound to lean toward the popular. To use a simple illustration, in the model's training datasets, the words "best" and "television series" are likely to appear in conjunction with "*The Sopranos*" more frequently than with a less popular Japanese series. In other words, the concentrated world that leans toward the popular and mainstream is a feature stemming from the fundamental technological paradigm underlying LLMs.[71]

Overall, these examples are consistent with the foregoing theoretical analysis: LLMs are likely to prioritize uniformity and convention over multiplicity and diversity. Despite the broad range of possible options, their outputs will plausibly reflect a concentrated worldview, projecting a thin slice of the world of thinkable thoughts—dominant narratives, blockbuster cultural products, and conventional choices—further reinforcing the popularity of the already popular.

Yet, why should we care about LLMs' inclination toward the standard? *The Sopranos*, after all, is a great series, Queen Victoria is undoubtedly a prominent nineteenth-century figure, and soy cheese can be used in a kosher cheeseburger. Should the mere dearth of plurality and diversity be a cause for concern? The next Section turns to this question.

C. The Social Costs

What, if any, are the social costs entailed in the apparent predisposition of LLMs toward a mainstream and narrow worldview? Ample multidisciplinary scholarship, ranging from sociology to cultural studies and deliberative democratic theory, establishes the significance of diversity and multiplicity. This scholarship highlights that exposure to various worldviews, languages, and cultures—global and local, popular and niche, national culture as well as "other" cultures—is both empowering at the individual level and a constructive factor for building social fabric.[72] Such

Duncan, Elke U. Weber & Eric J. Johnson, *When and Why Defaults Influence Decisions: A Meta-Analysis of Default Effects*, 3 BEHAV. PUB. POL'Y 159 (2019) (presenting a meta-analysis of dozens of studies on the influence of default choices); *cf.* Noga Blickstein Shchory & Michal S. Gal, *Voice Shoppers: From Information Gaps to Choice Gaps in Consumer Markets*, 88 BROOK. L. REV. 111 (2022) (describing how default choices of "voice shopper" algorithms can potentially harm consumers and markets).

[71] For a discussion of the next-word-prediction paradigm, see *supra* Section I.A.

[72] *See, e.g.*, C. EDWIN BAKER, MEDIA, MARKETS, AND DEMOCRACY 93–94 (2004); JOHN STUART MILL, ON LIBERTY 96–132 (1859) (stressing the importance of diversity and difference); Robert C. Post, *Democratic Constitutionalism and Cultural Heterogeneity*, 25 AUSTL. J. LEGAL PHIL. 185 (2000) (discussing the significance of cultural diversity for individual and group identity and examining its relations with the constitutional state); SEYLA

"diversity of exposure" raises awareness to different opinions, tastes, and perceptions; promotes tolerance and equality; and can serve as a buffer against extremism.[73]

Similarly, sociological research in the field of memory studies highlights the significance of multiplicity for collective memory.[74] This literature explains that collective memory—the ability of social groups to remember their joint past—is vital to forming group identity, constitutes a means of empowering minorities, and builds the individual's sense of self.[75] Importantly for our purpose, collective memory is not completely identical to historical accounts. The same historical event can play an entirely different role in the collective memory of different social groups.[76] Therefore, the ability of groups to form their collective memory depends on the availability of a multiplicity of voices and meanings.[77]

BENHABIB, THE CLAIMS OF CULTURE: EQUALITY AND DIVERSITY IN THE GLOBAL ERA (2002) (exploring the relations between cultural diversity and deliberative democracy); YOCHAI BENKLER, THE WEALTH OF NETWORKS: HOW SOCIAL PRODUCTION TRANSFORMS MARKETS AND FREEDOM 169 (2006) (explaining how "greater diversity of information, knowledge, and culture through which to understand the world" affects individual identity); *see also* Jürgen Habermas, *Three Normative Models of Democracy*, *in* DEMOCRACY AND DIFFERENCE: CONTESTING THE BOUNDARIES OF THE POLITICAL 21 (Seyla Benhabib ed., 1996); Cristina M. Rodriguez, *Language and Participation*, 94 CAL. L. REV. 687, 726–27 (2006) (explaining how language diversity nurtures individual identity and encourages participation); Miller McPherson, Lynn Smith-Lovin & James M. Cook, *Birds of a Feather: Homophily in Social Networks*, 27 ANN. REV. SOCIO. 415 (2001).

[73] *See* McPherson et al., *supra* note 72. For the term "diversity of exposure," see James G. Webster, *Diversity of Exposure*, *in* MEDIA DIVERSITY AND LOCALISM: MEANING AND METRICS 309 (Philip M. Napoli ed., 2007) (explaining that the term refers to the types of content actually consumed by people).

[74] For a primer on the concept of collective memory and its relations with the law, see Guy Pessach & Michal Shur-Ofry, *Intangibles and Collective Memory: The Role (and Rule) of Law*, 25 JERUSALEM REV. LEGAL STUD. 227 (2022). For a discussion of the significance of collective memory for groups and individuals, see, for example, Jeffrey K. Olick, *Collective Memory: The Two Cultures*, 17 SOCIO. THEORY 333, 333 (1999); Barbie Zelizer, *Reading the Past Against the Grain: The Shape of Memory Studies*, 12 CRITICAL STUD. MASS COMMC'N 214, 226–28 (1995); Jan Assmann, *Collective Memory and Cultural Identity*, 65 NEW GER. CRITIQUE 125, 126 (1995); W. James Booth, *The Work of Memory: Time, Identity, and Justice*, 75 SOC. RSCH. 237 (2008) (discussing the value of collective memory for the formation of individual identity).

[75] Booth, *supra* note 74.

[76] *See, e.g.*, Jeffrey K. Olick & Joyce Robbins, *Social Memory Studies: From "Collective Memory" to the Historical Sociology of Mnemonic Practices*, 24 ANN. REV. SOCIO. 105, 110 (1998) (discussing the multiplicity entailed in collective memory and the relations to historiography); Amos Funkenstein, *Collective Memory and Historical Consciousness*, 1 HIST. & MEMORY 5 (1989) (referring to the distinction between collective memory, which allows multiple narratives, and history). One example of an historical event that has a different role in the collective memory of different groups is the Battle of Alamo. *See* RICHARD R. FLORES, REMEMBERING THE ALAMO: MEMORY, MODERNITY, AND THE MASTER SYMBOL (2002).

[77] *See* FLORES, *supra* note 76.

This scholarship further instructs that a dearth of multiplicity and diversity can narrow our worldviews and might result in the exclusion of "others," those who do not conform to the standard and conventional.[78] Diversity and multiplicity are, therefore, crucial for social tolerance and stability and have a profound democratic significance.[79]

Importantly, diversity of exposure is not a static reflection of users' preferences. Rather, exposure to a multiplicity of content plays a role in *shaping* those preferences. Exposure to narrow, formulaic, and uniform content increases users' appetite for more formulaic content and decreases their demand for diverse content, and vice versa: Exposure to and engagement with a multiplicity of content positively affects the demand for diversity and can result in the growth of a "long tail" to the consumption curve.[80]

Against this analysis, the perils of diminishing diversity and narrowing worldviews in the age of LLMs transpire. LLMs are becoming a very dominant (perhaps *the* principal) prism through which people receive general information. They are already integrated into search engines, word processors, and a variety of consumer products that humans spend countless hours using.[81] Their outputs carry a meaning-making power. They provide a prism through which people view, and learn about, the world. As a result, their predispositions toward the mainstream and concentrated are likely to percolate and influence human perceptions: from the importance we attach to historical narratives, through the cultural products we select, to our choices of food.

This type of influence may well be elusive and almost invisible. Indeed, a user focused on a specific task, such as seeking historical information, a TV series recommendation, or a recipe, is unlikely to notice it. In fact, in the ordinary course of use, most people will likely settle for a single satisfactory output and will not even be aware of additional possibilities that these models can generate beyond the initial output. But in the long term, and in the aggregate, the narrow and concentrated prism that LLMs project will likely restrict human perceptions.

[78] McPherson et al., *supra* note 72, at 415–16 ("By interacting only with others who are like ourselves, anything that we experience . . . gets reinforced."). For a detailed discussion of this point, see Shur-Ofry, *Cultural Diversity*, *supra* note 56, at 207.

[79] *See* MILL, *supra* note 72, at 96–132; Post, *supra* note 72; BENHABIB, *supra* note 72; Habermas, *supra* note 72; Seyla Benhabib, *Models of Public Space: Hannah Arendt, the Liberal Tradition, and Jürgen Habermas*, *in* HABERMAS AND THE PUBLIC SPHERE 73, 82–83, 86 (Craig Calhoun ed., 1992); Rodriguez, *supra* note 72, at 726–27; Jack M. Balkin, *Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society*, 79 N.Y.U. L. REV. 1, 41 (2004) (discussing the significance of diversity to democratic culture).

[80] BAKER, *supra* note 72, at 30–31 (discussing the feedback loop between users' exposure to and demand for mass media products); ANDERSON, *supra* note 57 (describing a feedback loop between diverse supply and users' demand for "niche" products, including cultural products). For an elaborate discussion of this point, see Shur-Ofry, *Cultural Diversity supra* note 56.

[81] *See, e.g.*, Jaspreet Bindra, *Will ChatGPT Replace Google as Our Go-To Web Search Platform?*, MINT (Feb. 7, 2023, 4:33 PM), https://www.livemint.com/opinion/columns/will-chatgpt-replace-google-asour-go-to-web-search-platform-11671733523981.html [https://perma.cc/76JH-WL7A].

Yet are large language models different from other information structures, such as television, social media platforms, and search engines? Will the mediation of information through this new technology have a particularly powerful effect? The next Part unravels these questions by examining a combination of technological and design features that could make users particularly susceptible to the influence of LLMs.

II. USERS, MODELS, AND POWER RELATIONS

Ample studies in law, culture, and communication theory indicate that every medium conveying information necessarily reflects judgments regarding the meaning and importance of that information and, by so doing, inevitably imposes some normative prism on its users.[82] For example, communication theorist Edvin Baker has long observed that mass-media channels that expose viewers to easy-to-digest programs increase their appetite for more formulaic shows, at the expense of diverse and intricate content.[83] Likewise, research indicates that social media platforms often expose users to like-minded people rather than to diverse opinions—a phenomenon famously labeled "echo chambers"—resulting in increased polarization and extremism.[84] Search engines, too, influence their users' perceptions. Google, for example, describes its mission as "organiz[ing] the world's information and mak[ing] it universally accessible and useful."[85] Such organization necessarily entails a set of assumptions and priorities, coded in the search engine ranking algorithm, that reflect the judgment of its creators as to which results are more relevant than others. Interestingly, in the Google algorithm, too, popularity has a substantial weight in the ranking of search results. This implies that popular websites are prioritized in search results, which in turn may further increase

[82] *See, e.g.*, Niva Elkin-Koren, *Cyberlaw and Social Change: A Democratic Approach to Copyright Law in Cyberspac*e, 14 CARDOZO ARTS & ENT. L.J. 215 (1996) (noting that information structures impose upon their users the judgment of their creators); Eric Goldman, *Search Engine Bias and the Demise of Search Engine Utopianism*, 8 YALE J.L. & TECH. 188, 196 (2006) (explaining how search results reflect priorities and judgments of their creators); Michal Shur-Ofry, *Databases and Dynamism*, 44 U. MICH. J.L. REFORM 315 (2011) (discussing the meaning-making function of databases and their potential to influence users' perceptions); Michal Shur-Ofry & Guy Pessach, *Robotic Collective Memory*, 97 WASH. U. L. REV. 975, 987–89 (2020) (explaining how the mediation of historical events through algorithmic agents reflects human discretion and human choices).

[83] BAKER, *supra* note 72, at 30–31. For a discussion of this point, see also Guy Pessach, *Copyright Law as a Silencing Restriction on Noninfringing Materials: Unveiling the Scope of Copyright's Diversity Externalities*, 76 S. CAL. L. REV. 1067 (2003).

[84] *See, e.g.*, Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi & Michele Starnini, *The Echo Chamber Effect on Social Media*, PROC. NAT'L ACAD. SCI., Mar. 2021, at 1; YOCHAI BENKLER, ROBERT FARIS & HAL ROBERTS, NETWORK PROPAGANDA: MANIPULATION, DISINFORMATION, AND RADICALIZATION IN AMERICAN POLITICS (2018).

[85] *About Google*, GOOGLE, http://www.google.com/corporate/history.html [https://perma.cc/6SH3-27DS].

their popularity.[86] Algorithmic-based content recommendation systems similarly expose their users to a segment of content that is not likely to reflect true diversity.[87]

In the case of LLMs, however, a combination of design and technological features suggests that the influence on users' perceptions may be particularly powerful. The following paragraphs take a close look at these traits.

A. Encapsulation and Concealment

The output of LLMs, at least at the current technological phase, is often detached from the raw materials. LLMs, unlike search engines, seldom retrieve original sources.[88] Rather, the generation of new and condensed text out of a multitude of sources is at the core of this technology. Their user-driven output, together with the concise method of presentation—in a paragraph, a sentence, or a page—makes LLMs efficient and easy to use. Yet, these features also mask large parts of the relevant "world" in comparison to technologies that provide access to underlying materials and sources of information comprised by third parties. The single-paragraph output conveys an aura of authority and disguises the existence of myriad additional alternatives. As Michal Gal observed with respect to recommendation systems, "[a] user who is unaware of the algorithm's limitations, would likely not be aware of choices he has forgone."[89]

Consider, in comparison, a list of Google search results. Even if most of us do not get beyond the first page of search results, the mere knowledge that our query yielded, say, 50,000 results, makes us aware, in a subtle yet significant way, of the existence of a multitude of materials and additional relevant information. Likewise, the social media feed that triggered more than 12,000 replies to the "best television series" question alerted its viewers to multiple potential options, even if they merely browsed through the first replies.[90] Conversely, LLM technology simultaneously "encapsulates" and "conceals" the world, thereby creating an impression that the generated output is *the* answer.

[86] Goldman, *supra* note 82, at 193; *cf.* Lucas D. Introna & Helen Nissenbaum, *Shaping the Web: Why the Politics of Search Engines Matters*, 16 INFO. SOC'Y 169, 176, 181 (2000).

[87] *See* Jonathan Gingerich, *Is Spotify Bad for Democracy? Artificial Intelligence, Cultural Democracy, and Law*, 24 YALE J.L. & TECH. 227 (2022).

[88] Notably, this state of affairs might be somewhat changing, as some models are beginning to refer to source materials, yet such changes are gradual and far from all-encompassing. *See, e.g.*, Barry Schwartz, *Google Says Bard Won't Link to Sources Too Often*, SEARCH ENGINE ROUNDTABLE (Mar. 23, 2023, 7:51 AM), https://www.seroundtable.com/google-bard-wont-link-to-sources-too-often-35097.html [https://perma.cc/UJ4N-DE37] ("Bard, like some other standalone LLM experiences, is intended to generate original content and not replicate existing content at length.").

[89] Gal, *supra* note 9, at 74; *cf.* Leah Chan Grinvald & Ofer Tur-Sinai, *Smart Cars, Telematics and Repair*, 54 U. MICH. J.L. REFORM 283, 305 (2021) (observing how telematic systems in "smart cars" direct the users to specific repair options, while concealing others).

[90] *See supra* Section I.B.2.

B. Invisible Judgments

In many cases, the choices and priorities that underlie information structures are not easily observable by the users.[91] Yet, in the case of generative AI, this lack of transparency is particularly salient. As the previous discussion clarifies, the underlying technology does not allow the user to trace the model's "thinking process." This is not only because the ingredients that yielded the outcome—such as the training datasets, the hyperparameters, the human feedback, and the values assigned by human trainers—are largely imperceptible to users. It is also because the machine learning process includes "propagation," whereby the model receives feedback and adapts itself without explicit programming, and because the process leading to a particular final output does not lend itself to clear explanation, not even to the system's creators.[92]

C. "Enchantment," Anthropomorphism, and Trust

The asymmetrical power relations between LLMs and their users are buttressed by the human tendency to trust machine generated output, commonly referred to as "automation bias."[93] In our case, the phrasing of the output in a clear and often confident tone, using relevant jargon, creates an aura of authority and increases people's willingness to rely on it. This effect can be viewed as part of a broader "enchantment" phenomenon discussed in the literature, whereby people ascribe superhuman capacities to deep learning machines.[94] The ability to provide information in multiple fields and to generate new, high-quality output in a matter of seconds can bolster the Multivac Effect and the perception of these models as absolute arbiters, even when the user knows better. To illustrate, director Frank Pavich described how, after watching high-quality, AI-generated images of a film, which he knew did not exist, he nevertheless went searching for the film in databases: "I couldn't find anything because there was no film. There was no actor. There was no anything. These images were another A.I. creation. And I had known that right from the start. Yet still, I hoped that somehow it was real."[95]

In addition to the human deference to automated machines, LLMs that interact with users in a sociable, communicative way belong to a group of social robots—AI

[91] *See, e.g.*, GEOFFREY C. BOWKER & SUSAN LEIGH STAR, SORTING THINGS OUT: CLASSIFICATION AND ITS CONSEQUENCES 323 (1999) (explaining that structures of data often become "invisible").

[92] *See supra* Section I.A.

[93] *See, e.g.*, M.L. Cummings, *Automation Bias in Intelligent Time Critical Decision Support Systems*, AM. INST. AERONAUTICS & ASTRONAUTICS 1ST INTEL. SYS. TECH. CONF., Sept. 20, 2004, at 1 (discussing automated decision making and observing the human tendency not to search for additional or contradictory information in light of a machine generated solution that is "accepted as correct").

[94] *See* Alexander Campolo & Kate Crawford, *Enchanted Determinism: Power Without Responsibility in Artificial Intelligence*, 6 ENGAGING SCI., TECH., & SOC'Y 1 (2020).

[95] Frank Pavich, *This Film Does Not Exist*, N.Y. TIMES (Jan. 13, 2023), https://www.nytimes.com/interactive/2023/01/13/opinion/jodorowsky-dune-ai-tron.html?smid=url-share [https://perma.cc/2ZND-QV6S].

that engages with people in a sociable, cooperative, humanlike manner, demonstrating adaptability and learning skills.[96] Ample studies demonstrate that such social qualities of interaction elicit anthropomorphism—an inclination to attribute human qualities to the artificial intelligence.[97] The advanced skills of LLMs, the autonomous generation of new text, the vast "knowledge," the excellent communication skills, the ability to conduct a seemingly natural, humanlike conversation, and the ability to interact and to cooperate—all these qualities are bound to evoke an emotional response on the part of users. People, even sophisticated users, might treat LLMs as more than algorithmic tools.[98] Again, ChatGPT is a case in point. Users note that they feel an urge to use human pleasantries, such as "good morning," "please," and "thank you" when communicating with the robot.[99] Tech entrepreneur Aaron Levie even tweeted: "If you're not saying please and thank you in your ChatGPT conversations, then you've clearly never seen a sci-fi movie and good luck to you."[100] I myself have been anthropomorphizing throughout this Article, referring to the models' "knowledge," "thinking process," and "thinkable thoughts," and I still have an eerie sci-fi feeling each time an LLM asks me to confirm that I am not a robot.

[96] For the development of the concept of social robots, see, for example, Cynthia Breazeal, *Towards Sociable Robots*, 42 Robotics & Autonomous Sys. 167 (2003) (discussing the benefits of endowing robots with sociable skills); Kate Darling, *"Who's Johnny?" Anthropomorphic Framing in Human-Robot Interaction, Integration, and Policy*, *in* Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence 173 (Patrick Lin, Ryan Jenkins & Keith Abney eds., 2017) (examining ethical aspects related to social robots).

[97] For a discussion of the concept of anthropomorphism, see, for example, Kate Darling, *Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects*, *in* Robot Law 213, 213 (Ryan Calo, A. Michael Froomkin & Ian Kerr eds., 2016); Breazeal, *supra* note 96, at 168 (referring to the robot's learning capacity, creature-like behavior, and ability to communicate with, cooperate with, and learn from people, as the triggers for anthropomorphism); Shur-Ofry & Pessach, *supra* note 82, at 986–87.

[98] *Cf.* Darling, *supra* note 96, at 6; Matthias Scheutz, *The Inherent Dangers of Unidirectional Emotional Bonds Between Humans and Social Robots*, *in* Robot Ethics: The Ethical and Social Implications of Robotics 205, 213–14 (Patrick Lin, Keith Abney & George A. Bekey eds., 2012) (indicating that anthropomorphism does not disappear when people are aware of the underlying technology).

[99] To illustrate, see these posts on social media: Cameron Stow (@camerontstow), X (Jan. 23, 2023, 11:21 PM), https://x.com/camerontstow/status/1618101638782390272?s=46&mx=2 [https://perma.cc/N3NY-XPPF]; @smartfoodchef, X (Jan. 19, 2023, 2:03 AM), https://x.com/smartfoodchef/status/1615968231197401089?cxt=HHwWgoCzzeyiie0sAAA [https://perma.cc/PTT8-CU8R]; *Do You Ever Feel Sorry for ChatGPT?*, Reddit (Jan. 23, 2023), https://www.reddit.com/r/ChatGPT/comments/103gn3p/do_you_ever_feel_sorry_for_chatgpt/ [https://perma.cc/2UDB-FAGY].

[100] Aaron Levie (@levie), X (Dec. 6, 2022, 12:42 PM), https://x.com/levie/status/1600183992577187842?cxt=HHwWhICjzZ63_7QsAAAA [https://perma.cc/E8C6-B64L].

Taken together, these traits of LLMs are likely to evoke trust and reliance on the part of their users. To borrow Amisov's words, people will "ha[ve] faith" in their output. And as Asimov astutely observed: "That was what counted."[101]

D. AI Echo-Chambers

Finally, the power of LLMs to shape our universe of thinkable thoughts is expected to increase over time. The prevalent use of these models is likely to ignite feedback loops, whereby the texts generated by LLMs either percolate back into datasets or are intentionally used as "synthetic" data and affect the training of the next generation of LLMs. The result could be "*AI echo-chambers*," whereby AI feeds itself with its own thinkable thoughts.[102] And if datasets are flooded with LLMs' generated content, other speech will inevitably have less weight in the training materials.[103] These dynamics could amplify and reinforce the trends toward conformity at the expense of diversity and multiplicity. As OpenAI reflected in a technical report:

> As GPT-4 and AI systems like it are adopted more widely in domains central to knowledge discovery and learning, and as use data influences the world it is trained on, AI systems will have even greater potential to reinforce entire ideologies, worldviews, truths and untruths, and to cement them or lock them in, foreclosing future contestation, reflection, and improvement.[104]

***

The aggregation of the traits described here suggests that LLMs are likely to be very powerful in affecting their users. Preliminary empirical evidence supports this proposition.[105] The influence of these models on our perceptions—with a plausible

[101] ASIMOV, *supra* note 1, at 270.

[102] For discussions of different aspects of this phenomenon, see Eric Ulken, *Generative AI Brings Wrongness at Scale*, NIEMANLAB, https://www.niemanlab.org/2022/12/generative-ai-brings-wrongness-at-scale/ [https://perma.cc/94S7-NPQ4] (wondering whether the web will become "one big AI echo chamber"); Ethan Perez et al., Discovering Language Model Behaviors with Model-Written Evaluations 4 (Dec. 19, 2022) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2212.09251 [https://perma.cc/E6GR-JH24] ("[L]arger LMs are more likely to answer questions in ways that create echo chambers by repeating back a . . . user's preferred answer . . . ."); Ruibo Liu et al., Best Practices and Lessons Learned on Synthetic Data (Aug. 10, 2024) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2404.07503 [https://perma.cc/3YZ5-W9UV]; Rohan Taori & Tatsunori B. Hashimoto, *Data Feedback Loops: Model-Driven Amplification of Dataset Biases*, ICML'23: PROC. 40TH INT'L CONF. ON MACH. LEARNING 33883 (2023).

[103] *Cf.* HULL, *supra* note 42, at 19.

[104] OpenAI, *supra* note 5, at 49 (emphasis added).

[105] *See* Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson & Mor Naaman, Co-Writing with Opinionated Language Models Affects Users' Views (Feb. 1, 2023) (unpublished manuscript) (on file with arXiv), http://arxiv.org/pdf/2302.00560 [https://perma.cc/2D94-A9BV] (finding that a language-model-powered writing assistant that

shift from diversity and multiplicity toward uniformity, conformity, and a narrow worldview—is a real and imminent challenge.

Could personalization, namely, allowing users to customize LLMs to their tastes, provide the solution to the challenge of the narrowing world? Some LLM producers indeed enable users to customize the model to their needs.[106] However, it is doubtful whether personalization of LLMs could solve the diversity challenge. Even if users are able to customize the model's default outputs, they will still be exposed to a single, synthesized answer in response to their queries. For example, presenting a user with a Scandinavian television series in response to her "best television series" question might better align with her preferences, yet it would still mask the multiplicity of other options. In many cases, such personalization would echo existing views, thus making users less open to alternatives.[107] Moreover, as Jonathan Gingerich recently argued in the context of recommendation systems, personalization of content may offer "superficial diversity" but is unlikely to reflect "deep diversity" that could challenge extant users' perceptions.[108] The challenge, in other words, is not just to make *the model* more open to different perspectives but to maintain the awareness and openness of *users* to multiple perspectives.

Finally, can we expect a market solution to the diversity deficiency challenge? Is it likely that, if we just wait long enough, the LLM market will provide consumers with models whose outputs are more diverse, in a way that would obliviate policy intervention? The market of LLMs is still in stages of formation, and it is difficult to accurately predict its development. However, given the analysis above, there is solid reason to doubt that the market will shift itself toward multiplicity, plurality, and diversity in the default outputs of LLMs. Because the inclination toward mainstream and concentrated content results from fundamental traits of the technology,[109] overcoming the multiplicity deficit will require a conscious and deliberate effort on the part of the technology providers. Concomitantly, LLMs are expected to influence (and probably already influence) the tastes and worldviews of their users in a way that will further increase the (already substantial) demand for mainstream and uniform content, at the expense of variety and diversity.[110] It is highly doubtful that,

generates certain opinions more often than others impacts its users' opinions and writing).

[106] Alireza Salemi, Sheshera Mysore, Michael Bendersky & Hamed Zamani, LaMP: When Large Language Models Meet Personalization (June 5, 2024) (unpublished manuscript) (on file with arXiv), https://arxiv.org/abs/2304.11406 [https://perma.cc/UX24-W8CK]; *How Should AI Systems Behave, and Who Should Decide?*, OPENAI (Feb. 16, 2023), https://openai.com/blog/how-should-ai-systems-behave/ [https://perma.cc/DB8D-LQYJ] ("We believe that AI should be a useful tool for individual people, and thus customizable by each user . . . .").

[107] *See supra* text accompanying note 84 (discussing social media echo-chambers and the risk of extremism entailed in the lack of exposure to diverse views); *cf.* Erik Hermann, *Artificial Intelligence and Mass Personalization of Communication Content—An Ethical and Literacy Perspective*, 24 NEW MEDIA & SOC'Y 1258 (2021) (observing that AI-based content personalization can lead to selective exposure to specific content and limited content diversity, which may result in polarization, echo chambers, and filter bubbles, where individuals encounter content that reinforces their existing beliefs).

[108] Gingerich, *supra* note 87, at 271–72.

[109] *See supra* Section I.A.

[110] *See supra* note 80 and accompanying text. For a detailed analysis of the view that

absent intervention, AI manufacturers will have sufficient incentive to take diversity-increasing steps. Diversity and multiplicity, according to this analysis, can be seen as public goods that, in the age of LLMs, are subject to substantial market failure.

What, then, should be the legal-policy response to this intersection of LLMs and our universe of thinkable thoughts? The next Part takes a close look at the emerging AI governance schemes in the United States, alongside other prominent jurisdictions. Against this scrutiny, it introduces the concept of "multiplicity" and suggests that integrating this principle in AI governance would enable addressing the aforesaid challenges.

III. Integrating Multiplicity into AI Governance

A. Multiplicity and the Emerging AI Governance Landscape

The rapid developments in the field of AI in recent years yielded a rich discussion of AI governance schemes among scholars, industry players, and policymakers. Numerous proposals pertaining to the development and deployment of AI systems are being promoted in various jurisdictions, including, inter alia*,* the United States, the UK, Canada, and the European Union. As a general matter, the emerging trend among prominent jurisdictions is to base the regulatory frameworks in the field of AI on several core, high-level principles. I review these instruments in brief below against the overarching question that emerges from the previous analysis: Can core principles that are already prevalent in the field of information governance—such as transparency, explainability, or safety—sufficiently address the social challenges entailed in a multiplicity deficit, or, alternatively, should regulatory and governance frameworks explicitly recognize multiplicity and diversity as part of the high-order principles in the field of AI?

The United States has issued several guidelines and policies on AI during the past two years. All of them list central principles that should direct the design, use, and deployment of AI systems. The *Blueprint for an AI Bill of Rights*, released in 2022 by the White House Office for Science and Technology, set out five principles: safety, protection against discrimination, data privacy, notice and explanation, and providing human alternatives to algorithmic decision making.[111] The *AI Risk Management Framework*, released by the U.S. Department of Commerce-National Institute of Standards and Technology (NIST) shortly thereafter,[112] similarly detailed traits of trustworthy AI systems. The list includes "valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with harmful bias managed."[113] A presidential Executive Order

diversity cannot be left to the market, see Shur-Ofry, *Cultural Diversity*, *supra* note 56; *cf.* SARAH M. CORSE, NATIONALISM AND LITERATURE: THE POLITICS OF CULTURE IN CANADA AND THE UNITED STATES 58–61 (1997) (describing the Canadian approach that leaving culture to the free market is inconceivable).

[111] OFF. OF SCI. & TECH. POL'Y, BLUEPRINT FOR AN AI BILL OF RIGHTS: MAKING AUTOMATED SYSTEMS WORK FOR THE AMERICAN PEOPLE (2022).

[112] NAT'L INST. OF STANDARDS & TECH., ARTIFICIAL INTELLIGENCE RISK MANAGEMENT FRAMEWORK (AI RMF 1.0) (2023).

[113] *Id.* at 12–17.

on the *Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* from October 2023, revoked in January 2025, listed eight "guiding principles and priorities" for the development and use of AI, alongside detailed and binding provisions pertaining to their implementation. The Order's principles and priorities covered a wide range of topics that include, inter alia, safety, security, and reliability.[114]

The UK exhibits a similar approach, proposing to base the regulatory frameworks in the field of AI on five core principles, described in a command paper issued by the UK Office of Artificial Intelligence in 2023.[115] Those include "safety, security and robustness; appropriate transparency and explainability; fairness; accountability and governance; and contestability and redress."[116] Meanwhile, the proposed Canadian AI legislation aims to categorize AI systems according to their risk levels[117] and provides that those responsible for "high-impact" AI systems must establish measures to mitigate the risks of "harm or biased output" that could result from such systems. The current proposal leaves the definition of "high-impact" systems to future regulations.[118]

Without going into the subtleties of the differences between these schemes (which are important in themselves but are not essential to our discussion here), none of them includes reference to plurality, multiplicity, or diversity. Similarly, a new international Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law, adopted in September 2024 at the initiative of the European Commission, requires member states to implement in their artificial intelligence policies a series of principles, including, among others, preserving human dignity, transparency, accountability, equality, and privacy.[119] That list, too, does not include reference to diversity or multiplicity.

[114] Exec. Order No. 14,110, 88 Fed. Reg. 75,191 (Oct. 30, 2023). Additional principles are promoting innovation and competition in the field of AI, supporting American workers, advancing equity, safeguarding privacy and civil rights, protecting consumers, governing the use of AI by Federal agencies, and advancing international cooperation in the field. *Id.* The Order was revoked by the Trump administration on January 23, 2025. *See Removing Barriers to American Leadership in Artificial Intelligence*, WHITE HOUSE (Jan. 23, 2025), https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/ [https://perma.cc/PVK5-TMN6].

[115] OFFICE FOR ARTIFICIAL INTELLIGENCE, A PRO-INNOVATION APPROACH TO AI REGULATION, 2023, Cm. 815 (UK); *see also* DEPARTMENT FOR SCIENCE, INNOVATION & TECHNOLOGY, A PRO-INNOVATION APPROACH TO AI REGULATION: GOVERNMENT RESPONSE TO CONSULTATION, 2024, Cm. 1019 (UK).

[116] DEPARTMENT FOR SCIENCE, INNOVATION & TECHNOLOGY, A PRO-INNOVATION APPROACH TO AI REGULATION: GOVERNMENT RESPONSE TO CONSULTATION, 2024, Cm. 1019, at 43 (UK).

[117] Bill C-27, *An Act to Enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts*, 1st Sess, 44th Parl, 2022 [hereinafter Canadian AI and Data Act].

[118] *Id.*

[119] Council of Europe, Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, art. 6–13, Sept. 5, 2024, E.T.S. No. 225.

Finally, and importantly, the most detailed AI regulatory framework to date is the AI Act of the European Union, enacted in June 2024 after years of deliberation.[120] Similar to the Canadian approach, the Act attempts to distinguish between AI systems according to areas of activity and the level of risk involved. Systems defined as "high risk" are those that may create a significant risk of damage to health, safety, fundamental rights, the environment, and more.[121] Such high-risk systems have been subjected to stricter obligations concerning data governance, transparency, record keeping, security, and human oversight.[122] Like their counterparts in the United States, the UK, and Canada, these mandates do not refer to plurality, multiplicity, or diversity.

However, and important for our purpose, alongside these obligations, the AI Act's preamble refers to the *Ethics guidelines for trustworthy AI*, developed in 2019 by a group of experts on behalf of the European Commission.[123] The guidelines defined seven nonbinding ethical principles designed to ensure that the development and use of AI systems will promote the European approach of a "human-centric" AI.[124] Alongside principles such as technical safety, human agency and oversight, privacy, transparency and data governance, and technical robustness and safety, the list explicitly includes "diversity, non-discrimination and fairness."[125] The Act further elucidates that this phrase means that "AI systems are developed and used in a way that includes diverse actors and promotes equal access, gender equality and *cultural diversity* . . . ."[126] While the Act clarifies that the principles are not binding rules, it provides that they should be incorporated, "when possible," in the design and use of AI models, and further provides that the principles should serve as a basis for drafting the voluntary "codes of conduct" to be developed by the stakeholders in the AI field.[127]

On the one hand, then, the EU AI Act refrained from imposing "hard" mandatory obligations on AI providers with respect to diversity, settling instead on nonoperative statements in the Act's preamble and reference to voluntary codes of conduct. It is yet to be seen whether this extremely soft approach will have a practical effect on

[120] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June, 2024, Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), 2024 O.J. (L 1689) 1 [hereinafter the AI Act or the Act].

[121] *Id.* art. 6–7.

[122] *Id.* art. 8–15. For criticism of this distinction, see Michal Shur-Ofry, *A Networks-of-Networks Perspective on AI Policy*, NETWORK L. REV. (Mar. 4, 2024), https://www.networklawreview.org/shur-ofry-generative-ai/ [https://perma.cc/9HW6-BTYM].

[123] The AI Act, *supra* note 120, Preamble ¶ 27.

[124] *Id.*

[125] *Id.*

[126] *Id.* (emphasis added).

[127] *Id.* ("The application of those principles should be translated, when possible, in the design and use of AI models. They should in any case serve as a basis for the drafting of codes of conduct under this Regulation. All stakeholders, including industry, academia, civil society and standardisation organisations, are encouraged to take into account, as appropriate, the ethical principles for the development of voluntary best practices and standards.").

the AI landscape. Nevertheless, these provisions reflect an explicit recognition by legal policymakers that AI can adversely affect diversity, and that such potential effect is a cause for societal concern and should also concern relevant policymakers.

Should regulators explicitly incorporate diversity and multiplicity as core principles in their emerging AI governance schemes? Or are current principles in the aforesaid regulatory schemes—such as explainability, data security, or mitigating bias—sufficient to address the social concerns identified above? The following paragraphs indicate that while the extant principles might mitigate some of the social challenges expected to emerge with the proliferation of LLMs, they are unlikely to be sufficient to address the systemic risk of declining multiplicity.

Take, for example, explainability—one of the first principles in AI governance—which aims to tackle the black box nature of algorithmic decisions by imposing duties of explanation on AI developers.[128] Such explanation could identify biases and discrimination underlying algorithmic decision-making processes, which is extremely important when algorithms make decisions concerning individual rights like, for example, allocating credit score or risk-profiling. Explainability could also alert users to the human judgments and priorities embedded in LLMs' output and somewhat mitigate the Multivac Effect. In addition, having some information about the system's general design principles, and particularly the datasets of raw materials that underly LLMs, could give us an idea about the "universe" from which these models draw their output.

However, when the focus is on multiplicity of narratives and perceptions, explainability would be insufficient. Users who receive a reasonable answer to their inquiry about a historical event, a cultural product, or a recipe (to use our previous examples) are unlikely to seek explanations about underlying datasets and training principles, and, even if they did, the ability to trace and justify the origins of a specific output would be limited, at best.[129]

For related reasons, obligations to mitigate AI bias, such as those incorporated in some of the extant policy instruments, are unlikely to sufficiently address the problem of diminishing diversity. Treating any mainstream output (say, suggesting "Game of Thrones" in response to the "best series" question) as bias is impractical and unjustified. More broadly, the concept of bias in the context of AI regulation focuses on preventing discrimination against *individuals*.[130] Yet, the harms which this Article focuses on are largely *systemic*. These do not translate to immediate decisions affecting individuals. Rather, it is the incremental, cumulative effect of

[128] *See, e.g.*, Frank A. Pasquale, *Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society*, 78 OHIO ST. L.J. 1243, 1252 (2017) (describing explainability as providing information about the robot, including "to what has it been exposed, and how has this interplay between hardware, software, and the external environment resulted in present behavior").

[129] *See supra* notes 39–41, 91–92 and accompanying text.

[130] *See, e.g.*, Canadian AI and Data Act, *supra* note 117 § 5.1 ("[B]iased output [is] content that is generated, or a decision, recommendation or prediction that is made, by an artificial intelligence system and that adversely differentiates, directly or indirectly and without justification, *in relation to an individual* on one or more of the prohibited grounds of discrimination . . . .") (emphasis added).

LLMs' outputs over a large number of people across time that could yield undesirable and even grave societal consequences.[131]

Similarly, abiding by principles of data security and safety will reduce the risks of manipulation of LLMs' outputs and increase their reliability. Yet, it will not directly address the constraints that reliance on their outputs (however reliable) could pose for cultural diversity, collective memory, or world perceptions more generally. As the foregoing discussion demonstrates, general-purpose LLMs—however safe, explainable, privacy oriented, etc.—necessarily impose a constrained prism upon their users, which could result in a possible shift toward uniformity at the expense of diversity and multiplicity. Extant AI governance principles do not provide an easy fix to this problem.

Against this analysis, the need to introduce a principle of multiplicity into AI governance schemes becomes apparent. By "multiplicity" I mean exposing users, or at least alerting them, to the existence of multiple and diverse possible outputs, answers, narratives, and alternatives.[132] Adopting multiplicity as a governance principle will make users aware of the existence of different tastes and perceptions and allow them to glimpse at a universe that lies beyond the default output of LLMs. Moreover, the exposure to alternatives (such as more relevant historical figures, additional cultural products, or other nutritional options), and even the mere awareness that additional sources and options exist, could mitigate the Multivac Effect: decrease the authoritative power of LLMs and our inclination to automatically trust their default output. Instead, it will help people to view these models as they are: tools, rather than oracles. As a by-product, the mere exposure to various possibilities can also mitigate social biases—not by presenting an "objective" or "representative" reality, but merely by raising users' awareness to a host of potential narratives and views. Altogether, recognizing multiplicity as an AI governance principle and incorporating it as part of the high-order regulatory and ethical principles in the field will directly address the systemic risks of diversity deficits and narrowing perceptions in the age of LLMs.

Before we proceed, let us briefly address a possible concern: Will the incorporation of multiplicity in AI governance schemes be consistent with free speech principles? Recent scholarship has noted that generative AI can raise challenging questions from the perspective of free expression.[133] A comprehensive

[131] For the distinction between individual and systemic harms, see *supra* notes 15–17 and accompanying text.

[132] Notably, the notion of "multiplicity" is similar to the notions of "pluralism" and "diversity" (the latter is often used in the cultural context) but is broader and more general, and hence preferable for our purposes. I should also clarify that multiplicity as an AI governance principle is different from a concept that engineering Professor Ken Goldberg suggested in 2017, to describe a future where diverse groups of machines and humans will cooperate in a hybrid workforce, as opposed to the vision of "singularity" which implies the convergence of humans and robots. Ken Goldberg, *The Robot-Human Alliance*, WALL ST. J., (June 11, 2017, 4:39 PM), https://www.wsj.com/articles/the-robot-human-alliance-1497213576 [https://perma.cc/U5WK-HL3S].

[133] *See, e.g.*, Cass R. Sunstein, *Does Artificial Intelligence Have the Right to Freedom of Speech?*, NETWORK L. REV. (Feb. 28, 2024), https://www.networklawreview.org/sunstein-artificial-intelligence/ [https://perma.cc/X5MQ-CKC7]; Peter Henderson, Tatsunori

discussion of those questions require intricate distinctions between types of speakers and expressions against various theories of free expression and is beyond the scope of the analysis here. However, the focus on multiplicity does not pose the hardest questions in this context. First, introducing a principle of multiplicity into AI governance is essentially content-neutral. The idea is not to ban a particular speech or response by an LLM (and implicitly, possibly, by its AI providers), but rather the opposite. Acknowledging a principle of multiplicity would encourage a plethora of speech. By so doing, it would advance a diverse marketplace of ideas, which is, in fact, crucial from a free speech perspective.[134] In addition, as Jack Balkin observed with respect to social media platforms, under a theory of cultural democracy, users have rights to "participate in the forms of meaning-making that shape who they are and that help constitute them as individuals," as part of their own right to free expression, and such participation justifies imposing certain duties on information intermediaries.[135] The analysis in this Article therefore submits that embedding multiplicity as part of the governance frameworks that apply in the field of generative AI and imposing reasonable obligations on LLM providers to expose the users, or alert them, to the existence of diverse content, narratives, and cultural viewpoints would be consistent with free speech principles.[136]

Against this analysis, the next Section turns to examine how multiplicity can be integrated into AI governance. While this Article does not purport to offer a complete and exhaustive menu, the following sections explore two major avenues for such integration and examine possible legal frameworks that could accommodate this principle.

B. Implementation

1. Multiplicity-by-Design

One way of endorsing multiplicity in LLMs is "by design," namely by incorporating multiplicity-promoting features into the models' architecture. The idea of multiplicity-by-design draws on a broader understanding that certain values that society deems important can be embedded in and advanced through the architecture of technology. The most notable example to date is privacy-by-design, a principle which instructs that privacy considerations need to be taken into account during the engineering of technology and that the default choices of these architectures should

Hashimoto & Mark Lemley, *Where's the Liability in Harmful AI Speech?*, 3 J. FREE SPEECH L. 589 (2023).

[134] *See* Sunstein, *supra* note 133. For the "marketplace of ideas" justification for free speech, see, for example, Joseph Blocher, *Institutions in the Marketplace of Ideas*, 57 DUKE. L.J. 821 (2008).

[135] Jack M. Balkin, *Cultural Democracy and the First Amendment*, 110 NW. U. L. REV. 1053, 1061 (2016).

[136] *See id.*; *see also* Jack M. Balkin, *Information Fiduciaries and the First Amendment*, 49 U.C. DAVIS L. REV. 1183 (2016) (explaining that placing reasonable fiduciary obligations on social media platforms would not violate the First Amendment); Shur-Ofry & Pessach, *supra* note 82, at 995–96.

reflect privacy considerations.[137] Since its introduction not long ago, the notion of privacy-by-design gained considerable acceptance and was embraced by several regulators worldwide.[138]

Multiplicity-by-design is based on a similar notion: The architecture of LLMs should incorporate multiplicity-enhancing features that would direct users toward, or at least alert them to the existence of, diverse content, multiple worldviews, and various alternatives. A simple example already embedded in the architecture of some language models is the "regenerate" (previously "try again") button. This feature signals to the user that the initial one-paragraph default output provided in response to her prompt is not necessarily the only possible output and allows her to easily seek additional alternatives.

Another example for multiplicity-by-design concerns the phrasing of LLMs' output. When presented with a question that does not have a single correct answer, a tentative phrasing that explicitly acknowledges a spectrum of possibilities promotes multiplicity more than a curt and decisive response that presents the user with a "closed list." Let's return to our previous examples and consider some of the model's actual responses. A brisk and seemingly conclusive answer to a question about the most important nineteenth-century figures, such as "the three most important people who had lived during the nineteenth century are Napoleon Bonaparte, Queen Victoria, and Abraham Lincoln," buttresses the LLM's image as the ultimate authority. Consider, conversely, open-ended and provisional phrasings, such as "[i]t is difficult to say who the three most important people in the nineteenth century were, as it largely depends on one's perspective and what criteria are used to determine importance. Some notable figures . . . include . . . , others may argue that . . . ," or "[t]his is a highly subjective question, as different people will have different tastes in television shows. However, there have been a number of critically acclaimed television series . . . that many people consider to be among the best . . . [names of series] . . . . It's hard to pick a single show as it depends on various factors like genre, the individual's taste, mood, or what they're looking for . . . . It's always worth trying different shows to try to see what resonates with you personally." The latter outputs acknowledge that the issue involves discretion, alert the user to the existence of a range of possible views, and explicitly encourage them to explore further. The tentative tone leaves space for critical evaluations and reflections on the output and minimizes the Multivac Effect.[139]

[137] The development of this approach is attributed to Ann Cavoukian, Information and Privacy Commissioner of Ontario, Canada. *See* Ann Cavoukian, *Privacy by Design The 7 Foundational Principles: Implementation and Mapping of Fair Information Practices* (2011), https://privacy.ucsc.edu/resources/privacy-by-design---foundational-principles.pdf [https://perma.cc/9DGZ-GWTC]; *see also* Ira S. Rubinstein & Nathaniel Good, *Privacy by Design: A Counterfactual Analysis of Google and Facebook Privacy Incidents*, 28 BERKELEY TECH. L.J. 1333 (2013).

[138] *See* Ira S. Rubinstein, *Regulating Privacy by Design*, 26 BERKELEY TECH. L.J. 1409, 1410–11 (2012) (describing the regulatory acceptance of the principle). For implementation of privacy by design in the European Data Protection regime, see Council Regulation 2016/679, art. 25(1), 2016 O.J. (L 119) (EU).

[139] All these quotes are from actual responses of ChatGPT that were generated during the experimentations described in Section I.B. Copies on file with the author.

An additional feature, available in some of the LLMs, allows the user to calibrate one of the model's hyper-parameters called "temperature."[140] A higher temperature increases randomness and is likely to yield responses that are somewhat more diverse relative to the models' default, lower temperature.[141] However, there are some indications that a higher temperature may compromise the accuracy of the generated outputs, so that there might be some tradeoff between creativity and accuracy.[142] This implies that in order to achieve multiplicity while minimizing inaccuracies, users who calibrate the models' temperature would need to be mindful of these limitations and exercise more caution in relying on the output. This point highlights the more general need to increase AI literacy among users of LLMs, and I return to it shortly.[143]

The features discussed so far are relatively easy to implement. Additional, deeper, multiplicity-by-design steps may include a conscious, targeted effort to diversify the raw materials in training datasets by including materials from various cultures and languages.[144] Another feature could be linking the output of the models to relevant source materials. Such a design invites users to review these materials, independently assess them, and perhaps continue to explore.[145] An even more ambitious measure, which at the current state of technology seems aspirational, may entail a change in the "next-word-prediction paradigm," which grants inevitable weight to popularity in the generation of LLMs' outputs. Microsoft researchers recently maintained that this paradigm has inherent limitations, which manifest in GPT4's "lack of planning, working memory, ability to backtrack, and reasoning abilities."[146] Our analysis indicates that an additional limitation of the prevalent technological paradigm could be a concentrated, uniformity-inclined output, resulting in a multiplicity deficit.

The examples above are merely illustrative. Developing mechanisms for a robust multiplicity-by-design architecture obviously requires a combined multidisciplinary effort involving policymakers, stakeholders, computer scientists, social scientists,

[140] For the "temperature" parameter in LLMs, see, Hinton et al., *supra* note 47.

[141] *Id.; see also* Touvron et al., *supra* note 5, at 15 ("The temperature parameter also plays an important role for exploration, as a higher temperature enables us to sample more diverse outputs.").

[142] Touvron et al., s*upra* note 5, at 13 (describing such a phenomenon in Meta's LLM, Llama); _j, Comment to *Why the API Output Is Inconsistent Even After the Temperature Is Set to 0*, OPENAI DEV. F. (Aug. 25, 2023, 4:25 PM), https://community.openai.com/t/why-the-api-output-is-inconsistent-even-after-the-temperature-is-set-to-0/329541 [https://perma.cc/BC2A-NLKN] ("The sampling temperature, between 0 and 1. Higher values like 0.8 will make the output more random, while lower values like 0.2 will make it more focused and deterministic.").

[143] *See infra* notes 166–169 and accompanying text.

[144] *Cf.* Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite & Douwe Kiela, Measuring Data 2, 8–9 (unpublished manuscript) (on file with arXiv) (Feb. 13, 2023), https://arxiv.org/abs/2212.05129 [https://perma.cc/6HAE-5AV2] (suggesting measuring datasets across several measures, including, among other factors, their diversity).

[145] As the earlier discussion indicates, occasional references already exist, to a limited extent. *See* Schwartz, *supra* note 88.

[146] Bubeck et al., *supra* note 13, at 80; *see also supra* note 34 and accompanying text.

and engineers of LLM technologies. Given the growing effect these models are expected to have on our culture, collective memory, priorities, and world perceptions, this is a worthy endeavor.

2. Second (AI) Opinions

An additional way to promote multiplicity is by ensuring diversity in the LLM market to allow users access to several competing LLMs. Due to differences in underlying datasets, training processes, and output presentations, each of these sources is likely to reflect a (somewhat) different perception. Receiving "second AI opinions" would therefore allow users to compare various outputs and unravel additional "universes of thinkable thoughts." To quickly illustrate, consider again our culinary example.[147] Users could become aware of the option of a veggie burger if they have access to a second model whose default output is the vegan choice. A diversity of AI tools will further assist in diminishing users' enchantment and the perception of LLMs as "know-all" Multivacs. Finally, and parenthetically, although this Article focuses on cases where there's no single answer, "second AI opinions" could assist in detecting mistakes and falsehoods generated by AI in other cases where a single correct answer does exist.

The proposal to advance second AI opinions is somewhat reminiscent of recent scholarly proposals to promote oversight of algorithms through "AI oversight programs" that will review and audit AI decision-making.[148] In our case, however, promoting multiplicity requires no hierarchy between models. Rather, the mere prevalence of different AI tools will have a desirable effect.

Notwithstanding these advantages, the vision of a multiplicity of LLMs is far from simple. A growing body of research suggests that the field of AI itself is likely becoming more concentrated and less diverse. This scholarship indicates that the need to access enormous amounts of data, and the massive computing power required for developing deep-learning AI, may leave this arena in the hands of a small group of actors, most likely tech giants.[149] Some of the scholarship proposes regulatory interventions ranging from antitrust enforcement to the establishment of data-sharing

[147] *See supra* Section I.B.3.

[148] *See, e.g.*, Amitai Etzioni & Oren Etzioni, *Keeping AI Legal*, 19 VAND. J. ENT. & TECH. L. 133, 139 (2016).

[149] *See, e.g.*, Nur Ahmed, Muntasir Wahed & Neil C. Thompson, *The Growing Influence of Industry in AI Research*, 379 SCIENCE, 884 884–86 (2023) (discussing the concentration of AI in the hand of a small group of industry actors and observing that such concentration awards "a small number of technology firms an enormous amount of power over the direction of society"); Reza Shokri & Vitaly Shmatikov, *Privacy-Preserving Deep Learning*, CCS'15: PROC. 22D ACM SIGSAC CONF. ON COMPUT. COMMC'NS SEC. 1310, 1310 (2015) (suggesting that big tech-firms such as Facebook, Google, and Amazon have an advantage in AI research due to their access to massive data); Jonas Traub, Jorge-Arnulfo Quiané-Ruiz, Zoi Kaoudi & Volker Markl, Agora: Towards An Open Ecosystem for Democratizing Data Science & Artificial Intelligence (Sept. 6, 2019) (unpublished manuscript) (on file with arXiv), https://arxiv.org/abs/1909.03026v1 [https://perma.cc/54ZV-BHNL] (arguing that data sciences and AI are currently dominated by a small number of providers who can afford the massive investments required).

mandates obliging large companies that control data to provide access to that data to other entities.[150]

This discussion raises a more fundamental question concerning the involvement of the state in the emerging field of generative AI, not only as a regulator but also as an active stakeholder. Currently, prominent developers of LLMs are market-based corporations that operate in accordance with a set of market-based incentives. These incentives do not direct those stakeholders toward prioritizing multiplicity and diversity.[151] In past decades, the introduction of disruptive technologies in areas combining high barriers of entry with a potential to strongly influence public perceptions sometimes triggered state involvement as an actual technology provider. The ultimate example is public broadcasting, which many countries operate alongside private mass media channels. Indeed, studies indicate that the existence of powerful public broadcasting organizations (or lack thereof), and more generally, the extent of the state's investment in content, are significant factors that influence the level of diversity.[152] Therefore, if the state were to fund or provide "public LLMs," it would be more realistic to expect these models to prioritize multiplicity and diversity.[153]

The question of diversifying the AI landscape is certainly not limited to LLMs nor to the considerations of multiplicity and diversity but rather has implications for the entire AI field. Indeed, the need to foster competition in this field was set out as a priority in the 2023 presidential Executive Order.[154] While a complete review of the AI competition landscape and the extent of state involvement therein are beyond the scope of this study, the foregoing discussion contributes an additional angle to this debate by clarifying that diversifying the AI field to nurture (among other things) the availability of "second AI opinions" will also diversify the users' universe and mitigate the multiplicity deficit.[155]

[150] *See, e.g.*, Viktor Mayer-Schönberger & Thomas Ramge, *A Big Choice for Big Tech: Share Data or Suffer the Consequences*, 97 FOREIGN AFF. 48, 52–54 (2018) (discussing antitrust enforcement and suggesting a data sharing regime).

[151] *See supra* note 109 and accompanying text; *cf.* Kolt, *supra* note 15, at 1196–99 (explaining that commercial incentives in the AI industry direct the stakeholders toward a "steaming ahead" culture, while ignoring associated risks).

[152] *See* MICHÈLE LAMONT, MONEY, MORALS, AND MANNERS: THE CULTURE OF THE FRENCH AND AMERICAN UPPER-MIDDLE CLASS 140–45 (1992) (observing that a powerful public broadcasting system strengthens diversity and further maintaining that complete dependence on the market is unlikely to yield true diversity).

[153] *Cf.* Jennifer L. Schenker, *Can Europe Compete on Generative AI?*, INNOVATOR, (Apr. 23, 2023), https://theinnovator.news/can-europe-compete-on-generative-ai/ [https://perma.cc/8DQN-WTYF] (describing European initiatives of funding generative AI tailored to European priorities and values). Interestingly, Asimov's super-computer was also a state-owned entity. *See* ASIMOV, *supra* note 1.

[154] *See* Exec. Order No. 14,110, 88 Fed. Reg. 75,191 (Oct. 30, 2023).

[155] *Cf.* Mayer-Schönberger & Ramge, *supra* note 150, at 53 (mentioning that decentralization of data "would support diversity, innovation, and competition").

3. Legal-Regulatory Frameworks

What legal vehicles can accommodate multiplicity as a principle of AI governance? The following analysis reviews two potential routes and recommends a preferable path forward.

One ostensible alternative is to subject LLM providers to fiduciary duties and recognize multiplicity among those duties. This proposal builds on Jack Balkin's information fiduciary framework.[156] In his seminal research pertaining to social media platforms, Balkin maintained that digital organizations that collect large amounts of individual data should be subject to fiduciary duties due to the power they possess over users. That power results from a combination of trust—the willingness of users to trust these entities and believe they "will not betray" them—together with information asymmetries stemming from the fact that the collection and use of data about users is far from fully transparent to the users.[157]

In previous work with Guy Pessach, we proposed to extend the fiduciary framework and apply it to the use of algorithmic "memory agents"—human-like algorithms that mediate historical events and past experiences to the public. As we explained, the combination of trust and information asymmetries that underlies Balkin's information fiduciaries framework also subsists in the case of robots that mediate historical narratives.[158] The present analysis reveals a similar combination of trust, information asymmetries, and power relations in the case of LLMs. These models are likely to exert substantial influence over their users due to a series of traits: the distance between their output and the raw materials; the mode of output presentation; the invisibility of the processes, including the involvement of human judgments that influence the output; the ask-me-anything property that might trigger an enchantment effect; the communicative traits that trigger anthropomorphism and reliance; and the anticipated feedback loops that will further reinforce the models' point of view.[159] Similar to social media platforms, LLMs could become "forms of power that reshape and alter" us.[160]

This combination of trust, power, and information asymmetries in the relationships between LLMs and their users should give rise to fiduciary duties. The principle of multiplicity could be recognized as part of those duties. Seemingly, then,

[156] Balkin, *supra* note 136 (proposing the "information fiduciary" framework in response to rising privacy concerns in the digital age).

[157] *Id.* at 1185–86, 1223–32. For additional proposals to impose fiduciary duties on online platforms as a way to guard users' privacy, see, for example, ARI EZRA WALDMAN, PRIVACY AS TRUST: INFORMATION PRIVACY FOR AN INFORMATION AGE 85–92 (2018); *cf.* Woodrow Hartzog & Neil Richards, *Legislating Data Loyalty*, 97 NOTRE DAME L. REV. REFLECTION 356 (2022).

[158] Shur-Ofry & Pessach, *supra* note 82. For an additional proposal to apply the fiduciary framework to algorithmic decisions, see Brittany Swift, Note, *Artificial Constraints on Opportunity: Artificial Intelligence and Gender Discrimination in Automated Hiring Practices from an Information Fiduciary Perspective*, 28 B.U. J. SCI. & TECH. L. 215, 236–37 (2022).

[159] *See supra* Part II.

[160] Balkin, *supra* note 136, at 1211 (referring to social media platforms that collect data on their users).

the fiduciary structure provides a flexible and context-based framework for integrating the principle of multiplicity into AI governance, and, being based on judicial application of common-law principles, it does not necessitate explicit legislative intervention.[161] However, in our case, it also has a significant internal limitation. The preceding analysis clarifies that the social harms which are likely to result from a "multiplicity deficiency" are mainly *systemic*. In light of the systemic nature of the harm, individual users who receive valuable and reliable outputs to their distinct query on a recipe, or a name of a historical figure, are unlikely to initiate litigation. And even assuming they did, since the harm is cumulative and aggregate, rather than individual, establishing a breach of fiduciary duty in such distinct cases would be dubious.

The systemic nature of the problem, then, directs us toward an additional, preferable, legal-policy solution: incorporating a multiplicity principle in AI regulation and in AI ethical codes. As described earlier, regulators worldwide are currently debating how to address various algorithmic challenges and which standards should be imposed on AI providers.[162] Some of the regulatory instruments explicitly state that their aim is not only mitigating individual harms but also risks to "society as a whole."[163] A principle of multiplicity can address risks that could lead to systemic societal harms and should therefore be incorporated into the existing high-order principles that underpin AI regulatory frameworks.

Notably, incorporating multiplicity into AI governance frameworks does not necessarily imply imposing mandatory obligations on AI providers. A combination of "soft law" and regulatory incentives may be sufficient for this purpose, and the precise measures adopted would depend to a large extent on developments in the AI market. Yet, the discussion above indicates that leaving the issue to the market alone is likely to result in a serious multiplicity deficit. Therefore, even if the principle is anchored as a "soft" ethical or voluntary standard, it is important that policymakers provide incentives that will encourage the stakeholders to adopt and implement it, for example, through regulation that provides legal or business advantages to firms that meet this standard.

Importantly in this context, several industry stakeholders have declared that their endeavors in the AI field will abide by certain ethical standards. Google's AI principles, specifically, state the company's commitment to "socially beneficial" AI and include reference to diversity in the creation of the training datasets, in tagging the information therein, and in assessing their safety.[164] Concomitantly, according to reports, there are efforts to draft voluntary codes of conduct in the field of AI that

[161] *Cf.* Shur-Ofry & Pessach, *supra* note 82.

[162] *See supra* Section III.A.

[163] *See* AI Act, *supra* note 120, Preamble ¶ 130, art. 3(65) (reference to "systemic risks"); *cf.* GOV'T OF CAN., THE ARTIFICIAL INTELLIGENCE AND DATA ACT (AIDA) – COMPANION DOCUMENT, https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document [https://perma.cc/9HEL-HTLE] (referring to AI's "potential to cause harm to society").

[164] *See* GOOGLE, AI PRINCIPLES PROGRESS UPDATE 2023, (2023), https://ai.google/static/documents/ai-principles-2023-progress-update.pdf [https://perma.cc/9TQV-39CD].

would be mutually adopted by EU and U.S. regulators.[165] Therefore, explicit reference to and incorporation of a multiplicity principle in AI governance schemes could percolate to the industry and contribute to the emergence of a de facto ethical standard that would be adopted by relevant stakeholders, even in the absence of mandatory obligations.

Finally, legal, ethical, and regulatory solutions alone may not be sufficient to resolve the challenges of increased conformity and narrowing worldviews. This Article's analysis clarifies that the challenges of diminishing diversity are inextricably related both to the properties of the new technology, as well as to the markets in which AI providers operate. Moreover, ample literature suggests that diversity is a multi-causal phenomenon that cannot be resolved through a single regulatory intervention.[166] Therefore, the effort to maintain multiplicity in the era of LLMs should explore and advance additional measures, beyond the legal-regulatory solutions. One such important step could be encouraging *AI literacy* among LLM users.[167] AI literacy implies, in our context, that users should have a basic understanding of how LLMs work; how their output is affected by human discretion, dataset availability, and popularity; and why they can have a substantial influence on our worldviews. Attaining AI literacy would empower users, highlight "the[ir] own capacity to decide,"[168] and encourage them to seek additional information. To paraphrase Asimov (one last time), AI literacy would help us critically evaluate the responses we receive from LLMs and to realize that their outputs are not always "the best available," and are certainly not all what counts.[169]

CONCLUSION

Society has just begun its acquaintance with generative AI. The exploration of LLMs, their enormous potential alongside their social implications, is in a nascent stage. The challenges entailed are still to transpire. The analysis in this Article demonstrates that these challenges will not be confined to questions of misinformation, errors, or misuse. LLMs could emerge as powerful tools that shape us in subtle but deeper ways. In time, they might restrict the prism through which we view the world, affect cultural diversity and collective memory, and narrow our universe of thinkable thoughts. As the use of LLMs becomes ubiquitous, the

[165] *See, e.g.*, Natasha Lomas, *EU and US Lawmakers Move to Draft AI Code of Conduct Fast*, TECHCRUNCH, (May 31, 2023, 10:14 AM), https://techcrunch.com/2023/05/31/ai-code-of-conduct-us-eu-ttc/ [https://perma.cc/7QS3-UU84]; Camille Ford & Carisa Nietsche, *US-EU AI Code of Conduct: First Step Towards Transatlantic Pillar of Global AI Governance?*, EURACTIV, (July 27, 2023, 11:05 AM), https://www.euractiv.com/section/artificial-intelligence/opinion/us-eu-ai-code-of-conduct-first-step-towards-transatlantic-pillar-of-global-ai-governance/ [https://perma.cc/DT7F-MMLG].

[166] *See* Shur-Ofry, *Cultural Diversity*, *supra* note 56.

[167] For proposals to promote AI literacy as a way to mitigate challenges emerging in the AI field, see, for example, Hermann, *supra* note 107, at 1270–71 (arguing that AI literacy could empower individuals and reduce the challenges entailed in AI-driven mass personalization).

[168] *Id.* at 1270.

[169] ASIMOV, *supra* note 1.

influence of these prisms will increase as well. People, tastes, and events, which LLMs depict as central and important, will become even more central, while those remaining outside the models' judgments and thinkable thoughts will be relegated to the fringes. Ignoring these challenges could cause substantial social harm because what is at stake "is [ou]r own selves."[170]

Currently, most AI governance frameworks do not propose a satisfactory solution to the concerns of diminishing diversity and narrowing worldviews. Introducing multiplicity into AI discourse and incorporating this principle into AI governance will directly address these challenges and allow legal policy to keep pace with developments in the field of AI. Maintaining a diversity of narratives, content, and perceptions in our intersection with artificial intelligence is a multi-faceted challenge, and a multiplicity principle alone may not provide a magical solution. Yet, integrating it in AI governance schemes could promote the development of technological and legal frameworks that would significantly advance this social goal. As our relations with generative AI are entering a new phase, recognizing multiplicity as an AI governance principle will allow us to benefit from these technologies without sacrificing the complexities and intricacies of the human experience.

[170] Balkin, *supra* note 136, at 1211 (referring to social media platforms).