Risk and Harm

Unpacking Ideologies in the AI Discourse

Gabriele Ferri
Eindhoven University of Technology, Eindhoven,
Netherlands
g.ferri@tue.nl

Inte Gloerich
Utrecht University, Utrecht, The Netherlands,
i.gloerich@uu.nl; Institute of Network Cultures,
Amsterdam University of Applied Sciences, Amsterdam,
The Netherlands

ABSTRACT

We examine the ideological differences in the debate surrounding large language models (LLMs) and AI regulation, focusing on the contrasting positions of the Future of Life Institute (FLI) and the Distributed AI Research (DAIR) institute. The study employs a humanistic HCI methodology, applying narrative theory to HCI-related topics and analyzing the political differences between FLI and DAIR, as they are brought to bear on research on LLMs. Two conceptual lenses, "existential risk" and "ongoing harm," are applied to reveal differing perspectives on AI's societal and cultural significance. Adopting a longtermist perspective, FLI prioritizes preventing existential risks, whereas DAIR emphasizes addressing ongoing harm and human rights violations. The analysis further discusses these organizations' stances on risk priorities, AI regulation, and attribution of responsibility, ultimately revealing the diverse ideological underpinnings of the AI and LLMs debate. Our analysis highlights the need for more studies of longtermism's impact on vulnerable populations, and we urge HCI researchers to consider the subtle yet significant differences in the discourse on LLMs.

CCS CONCEPTS

• **Computing methodologies** → Artificial intelligence; Philosophical/theoretical foundations of artificial intelligence; • **Human-centered computing** → Human computer interaction (HCI); • **Applied computing** → Law, social and behavioral sciences; Arts and humanities; • **Social and professional topics** → Computing / technology policy.

KEYWORDS

AI, Large Language Models, Politics, Ideology, Longtermism, Human Rights

ACM Reference Format:
Gabriele Ferri and Inte Gloerich. 2023. Risk and Harm: Unpacking Ideologies in the AI Discourse. In *ACM conference on Conversational User Interfaces (CUI '23), July 19–21, 2023, Eindhoven, Netherlands.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3571884.3603751

1 INTRODUCTION

As large language models (LLMs) and AI applications continue to surge forward, both excitement and concern ripple through the public discourse. The public release of GPT3, GPT4, and ChatGPT spurred frantic competition among tech giants, with Microsoft announcing a conversational version of its Bing search engine [35] and Google following soon thereafter [22, 23]. At the same time, the academic [17, 31, 38, 43] and popular discourses [24, 25, 30, 40] around these technologies oscillate between amazement and moral panic [14, 20, 28, 34, 46]. Higher education is not immune, with heated discussions on the future of academic standards, peer review, and students' assessments [15, 27, 42, 47].

A critical juncture unfolded between March 29th and 31st, 2023. On March 29th, the Future of Life Institute (FLI) released[1] an open letter [19] advocating for a six-month moratorium on "training AI systems more powerful than GPT-4." On the same day, Time magazine published an op-ed by Eliezer Yudkowsky warning that building a "too-powerful AI" could destroy Earth's biological life [49]. By March 30th, Italy's data protection authority temporarily[2] banned ChatGPT, investigating whether it violated the GDPR by using conversations as training data and exposed minors to age-inappropriate content [44]. Finally, on March 31st, the Distributed Artificial Intelligence Research Institute (DAIR) published a response to FLI's letter, advocating for transparent regulation to protect vulnerable populations [21].

All these public initiatives share the goal of implementing stricter regulations on the development of LLMs and related AI technologies. However, despite this shared objective, they are rooted in fundamentally different ideologies, which often go unacknowledged. Amidst these intertwined discussions, our objective is to dissect the ideological underpinnings of the positions in the debate on LLMs and AI. By analyzing the back-and-forth between FLI and DAIR [19, 21], we tease out two competing political visions shaping the discourse around AI governance, and we aim to help researchers to orient themselves in this rapidly-evolving debate.

2 METHODOLOGY

This work is situated in Humanistic HCI [3], an interdisciplinary perspective combining methodologies and concepts from the humanities with the application domain of Human-Computer Interaction. We focus our analysis on the FLI and DAIR letters, and, secondarily, we also consider other webpages hyperlinked to from

CUI '23, July 19–21, 2023, Eindhoven, Netherlands
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0014-9/23/07.
https://doi.org/10.1145/3571884.3603751

[1] The FLI open letter was first circulated amongst selected recipients from March 22nd, 2023, but it remained under embargo until March 29th. In the following weeks, other content (a FAQ, some additional documents, . . .) was added to the webpage. In our analysis, we consider the letter's text as it was first published on March 29th.

[2] The service was restored on April 28th [26].

the primary texts[3]. We apply close reading, a method from narrative theory [12], to produce a detailed and technically informed examination of the FLI and DAIR's letters [19, 21]. These texts are of interest to CUI and HCI as they exemplify the current discourse around LLMs and to the humanities for their socio-cultural significance.

We followed the same analytic procedure for both texts. First, we archived the first public version of each letter [19, 21]. We examined their cultural and academic context of production by scanning other recent publications of the same authors and recent texts referring to them. From this, we identified two conceptual lenses ("existential risk" and "ongoing harm") that we used to orient our analysis. Then, we segmented the letters into sequences and open-coded them [1, 11]. This yielded three thematic clusters ("risk priorities," "AI Regulation," and "attribution of responsibility"). We described each cluster through the lenses of existential risk and ongoing harm. In conclusion, we critically interpret the two texts and their ideological perspectives.

3 CONCEPTUAL LENSES

We introduce two conceptual lenses, existential risk and ongoing harm, shaped by the cultural and academic context of the FLI and DAIR letters.

3.1 Existential Risk

The "lens of existential risk" expresses a longtermist perspective[4] [32, 39], which prioritizes maximizing humanity's potential in an astronomically long timeframe. The reason why we turn to longtermism for this conceptual lens is twofold. First, it is evoked in the DAIR letter, accusing FLI of addressing points that are "the focus of a dangerous ideology called longtermism." Additionally, we confirmed through the openphilanthropy.org website that FLI received annual grants under the "longtermism" category.

To understand existential risk from a longtermist view, we must first delve into its core ethical tenets. Longtermism [32, 39] is a philosophical perspective concerned with humanity's "full potential," determined with a quantitative and utilitarian formula. Such potential is calculated by estimating the number of people born throughout humanity's existence and taking that as an absolute parameter: the more humans expand and multiply, the higher humanity's achieved potential is. Furthermore, the speculative models underlying longtermist moral calculus assume that humanity will eventually colonize space, evolve into a posthuman technologically-augmented race, and produce sentient beings living in computer simulations [7, 8, 10]. When the staggering amount of yet-to-be-born spacefaring posthuman lives is compared with the humans living now on Earth, longtermists believe that we have a moral duty to the former by their sheer number.

This belief leads to morally contentious conclusions, such as Bostrom's foundational work [7] advocating for maximizing economic productivity without considering climate change – which, in a longtermist perspective, is a minor setback worth the losses it produces. Furthermore, Beckstead writes: *"Saving lives in poor countries may have significantly smaller ripple effects than saving and improving lives in rich countries. Why? Richer countries have substantially more innovation, and their workers are much more economically productive"* [4]. Longtermists' position on AI follows from their ethical tenets and can be traced back to Bostrom's work [7–9]. AI is an essential element in the road towards this longtermist future, but one that needs to be managed well. Bostrom recognizes that a loyal AI system could hasten human advancement but contends that disloyal AIs are existential risks. Longtermist moral calculus deems ethically wrong anything that lowers humanity's potential; therefore, AI research is viewed with extreme caution.

Adams et al. [2] critique this perspective, stating that *"Longtermists encourage us to deprioritize the needs of existing people, whose numbers pale in comparison to those of future generations. This focus on an abstract future often results in overlooking the concrete ways in which that future will materialize."* In the context of our analysis, the "lens of existential risk" teases out elements that, in a longtermist view, increase the risk of a hypothetical rogue AI. When analyzing a text through this lens, we highlight all elements that longtermists would deem ethically wrong.

3.2 Ongoing Harm

The "lens of ongoing harm" addresses human rights-based approaches to AI [33] and, to illustrate its significance for our analysis, we refer to a paper by Prabhakaran, Mitchell, Gebru, and Gabriel [41], which is notable because two authors (Mitchell and Gebru) are also co-signatories of the DAIR letter. They argue that implementing a human rights-based approach to AI allows developers to align systems with various socio-cultural contexts. It would also enable them to acknowledge the responsibilities and duties of different actors, including AI systems, towards individuals and create frameworks that involve historically marginalized communities in essential AI system decisions. Prabhakaran et al. [41] argue that addressing AI research with a human rights-based approach helps forge a stronger connection between AI models, the socio-technical systems they operate in, and the ongoing harm caused by AI. The authors suggest that this approach can support a reorientation away from formal principles and towards human welfare and a person's capacity to flourish. Additionally, the authors suggest that human rights frameworks offer well-established principles centered on human vulnerabilities and needs, promoting greater accountability in AI systems.

Historically, human rights and values have played a vital role in analyzing and designing digital artifacts. Value Sensitive Design (VSD) is a well-known perspective in HCI that offers a *"theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process"* [18]. It has evolved over the past two decades, with attempts to connect VSD to legal frameworks such as the European Convention on Human Rights [29] and the Convention on the Rights of Persons with Disabilities [13].

When we apply the lens of ongoing harm in our analysis, we bring to the forefront contexts where people experience injustice, human rights violations, and other damages. It helps us to uncover and address the issues that AI technologies create for vulnerable

[3] It is also significant that the FLI letter cites a famous paper by the DAIR authors [5], and that the DAIR letter is a direct and critical response to FLI.

[4] As a specific philosophical term, "longtermism" does not correspond to the general notions of "planning for the long-term" or "caring for future generations." See Moreno's recent work unpacking its "anti-presentist" ideology [37].

Table 1: Sequences in the FLI letter

Sequence title	Summary
The risks of advanced AI	The FLI letter expresses concerns about the potential risks associated with advanced AI, warning that AI can surpass human intelligence, resulting in unprecedented societal change.
Crucial decisions on AI	It highlights that humanity faces crucial decisions on AI development, focusing on the ethical implications of developing non-human minds that could eventually replace humans.
Call for pausing and refocusing AI research	It calls for a pause of AI research, urging to assess the potential risks and benefits of AI development and to redirect efforts towards ensuring AI safety, value alignment, and long-term societal flourishing.
Proposals for AI governance	It proposes that AI governance should involve AI developers, independent experts, and policymakers, ensuring that AI policy follows a precautionary approach.

Table 2: Sequences in the DAIR letter

Sequence title	Summary
Critique of the FLI letter's focus on hypothetical risks	The DAIR letter critiques FLI's focus on hypothetical risks, arguing that it diverts attention from present-day problems and that FLI's emphasis on potential future existential threats is disproportionate to the urgency of addressing current issues.
Call for transparency and accountability in AI	It calls for accountability, emphasizing that AI developers and companies should be held responsible for the societal consequences of their technologies, and that AI systems should be designed to prioritize human rights.
Emphasis on addressing current issues and social equity	It underscores the need to address ongoing social inequities exacerbated by AI systems, focusing on the harm caused by worker exploitation, data theft, and concentration of power. It advocates for inclusive AI governance.

populations by focusing on the potential harm inflicted on individuals. By employing this lens as a conceptual tool, we reveal issues and opportunities to incorporate human rights principles into AI design and investigate how AI can support human rights practices.

4 ANALYSIS

We present our close reading of the FLI and DAIR letters. We identified sequences and actors expressed in the texts and analyzed how their arguments were constructed. Through open coding, we identified three thematic clusters: "Risk Priorities," "AI Regulation," and "Attribution of Responsibility."

4.1 Sequences

We begin by identifying the significant sequences and actors present in the texts.

4.2 Risk Priorities

We examine the elements that are indicated by FLI and DAIR as severe dangers brought forward by AI. By doing so, we aim to make explicit the priorities expressed in the two letters. The "lens of existential risk" brings to the forefront FLI's concern about superhuman AIs, coherently with longtermist concerns. On the other hand, DAIR's concerns – which are not projected in the future but are grounded in the here-and-now – come across clearly through the "lens of ongoing harm."

FLI's rhetoric is emphatic, with passages like *"Should we develop non-human minds that might eventually outnumber, outsmart, obsolete and replace us?"* and *"out-of-control race to develop and deploy ever more powerful digital minds that no one – not even their creators – can understand, predict, or reliably control."* The risk is presented as overwhelming ("non-human minds that might eventually outnumber us") and uncontrollable ("no one [. . .] can understand [AI]") for the whole humanity ("obsolete and replace us"). FLI paints a picture of a deadly fight between humanity and AI, set in an unspecified future ("eventually").

Conversely, DAIR's text prioritizes addressing present concerns about developing and deploying AI technologies. Their rhetorical structure is based on hyperlinking key texts [6, 16, 36, 45, 48, 50] to their argument. In what follows, we synthesize the documents they refer to. DAIR mentions worker exploitation in AI-driven industries, where underpaid gig workers, such as data laborers and content moderators, experience precarious labor conditions and are often recruited from impoverished populations [6, 48]. Additionally, DAIR mentions intellectual property theft through AI-generated content [36] (e.g., Stability AI and Midjourney have been accused of using others' intellectual property without permission). DAIR's letter also raises concerns about the opacity and lack of transparency in search systems [45], which can lead to transparency issues and potentially negative consequences for information verification. Lastly, it emphasizes the risk of automation exacerbating existing inequalities, and AI systems can reinforce and perpetuate these injustices [16, 50].

4.3 AI Regulation

A second cluster we found in our analysis pertains to AI regulation and governance proposals presented in the two letters. We tease out the authors' conceptions of humanity and society in relation to AI and their implicit visions of AI regulation.

FLI and DAIR's different conceptualizations of humanity are evident in the letters. On one side, FLI adopts rhetorical structures constructing humankind as a monolithic collective actor, such as: *"Should we let machines flood our information channels with propaganda and untruth? Should we risk loss of control of our civilization?"* For FLI, humanity is defined in opposition to non-human minds, with a construction of "otherness" that implies an almost alien inscrutability. Vice versa, DAIR highlights the diversity of actors developing AI technologies, deploying them, or being affected by them, often negatively. As discussed above, DAIR strategically uses hyperlinks to point at marginalized communities, tech and gig workers, and creative professionals at risk of intellectual property theft.

The differing approaches of each text toward AI governance and decision-making further reflect the distinction between monolithic and nuanced views of humanity. DAIR emphasizes inclusive processes that consider the perspectives of marginalized and affected communities. They write: *"Those most impacted by AI systems, the immigrants subjected to "digital border walls," the women being forced to wear specific clothing, the workers experiencing PTSD while filtering outputs of generative systems, the artists seeing their work stolen for corporate profit, and the gig workers struggling to pay their bills should have a say in this conversation."* In contrast, FLI implies a top-down approach to governance by suggesting that *"AI developers must work with policymakers to dramatically accelerate the development of robust AI governance systems."*

In FLI's view, concerns of a smaller scale, such as the experience of underserved minorities, do not register through the lens of existential risk and are given a lower priority. Vice versa, they are outstanding when observed through the lens of ongoing harm and are central to DAIR's text. Furthermore, the two letters reveal opposing views on society and technology. Through the lens of existential risk, FLI's tech-driven stance is evident when they write, *"we can now enjoy an AI summer in which we [. . .] give society a chance to adapt."* Society adapting to AI also refers to longtermism's ambition for posthuman enhancement [10]. From the opposite position, DAIR's text argues that technology should adapt to society to protect human rights: *"we do not agree that our role is to adjust to the priorities of a few privileged individuals and what they decide to build and proliferate. We should be building machines that work for us, instead of adapting society to be machine readable and writable."* DAIR's text reinforces its plural understanding of humanity, emphasizing the different underlying assumptions about the role of technology in shaping society.

4.4 Attribution of Responsibility

Finally, we present the cluster where FLI and DAIR touch upon accountability and AI technologies. In this section, we examine different views on the responsibilities of developing and running AI-based tools.

DAIR underscores the responsibility of for-profit corporations and startups for driving AI development. They write, *"The onus of creating tools that are safe to use should be on the companies that build and deploy generative systems, which means that builders of these systems should be made accountable for the outputs produced by their products."* Analyzing this passage through the lens of ongoing harm reveals DAIR's portrayal of AI as a human-made technology, with developers held responsible for its consequences, rather than viewing AI as mysterious, emergent black boxes, as FLI seems to imply.

In contrast, FLI's anthropomorphization of AI systems ("artificial digital minds" and "systems with human-competitive intelligence") reflects longtermist concerns about possible existential risks from misaligned AI systems. By depicting them as autonomous entities capable of independent thought, FLI suggests that accountability is distributed among the AI, policymakers who fail to regulate their development, and the developers themselves. In other words, responsibility for any harm caused by AI systems is not solely placed on their creators.

This difference in attributing responsibility also implies distinct approaches to addressing the challenges posed by AI, stemming from their divergent concerns. While both FLI and DAIR advocate for more regulation, they do so from significantly different perspectives. FLI, concerned about AI as a potential existential risk if not developed with extreme caution, emphasizes the need for robust AI alignment research and the development of fail-safe mechanisms to ensure AI systems remain aligned with human values. This focus on AI systems' potential autonomy and agency reflects their belief that AI poses a risk if not developed with utmost care. On the other hand, DAIR is concerned with avoiding injustices and ongoing harm and calls for more immediate actions such as regulation, transparency, and inclusiveness in AI development processes. Their focus on the responsibility of the broader system highlights their commitment to addressing the challenges posed by AI in the context of present-day issues and inequalities.

5 CONCLUSIONS: URGENT TAKEAWAYS FOR HCI

We have teased out some ideological underpinnings in the discourse on LLMs and AI. The dichotomy between the lenses of existential risk and ongoing harm is the crux of the disagreement between key voices[5] in AI regulation, governance, and societal impact. FLI adopts a longtermist perspective, focusing on potential existential risks posed by AI in the distant future. This perspective, while highlighting potential future threats, diverts attention and resources from addressing ongoing harms and violations of human rights,

[5] While we focused on FLI and DAIR, also the other two examples mentioned in the Introduction could be read along the same lines, with Yudkowsky [49] taking a longtermist position and the Italian SA authority [26, 44] implicitly adopting a human rights-inspired stance that is closer to DAIR.

thereby exacerbating social inequities. In stark contrast, DAIR emphasizes the immediate harms of AI technologies. They advocate for transparency and accountability, focusing on the urgency to address present-day challenges that include human rights, inclusiveness, and ongoing harm. This focus recognizes the importance of inclusive, participatory processes that account for the voices of marginalized and affected communities.

HCI researchers cannot lose sight of the fact that adopting AI is not merely a question of which processes to automate but a deliberate choice with far-reaching consequences. HCI's role extends beyond simply exploring the technological frontier; it also entails asking critical questions about ethics. The call for AI regulations is fraught with ideological complexities: FLI and DAIR's perspectives are not equivalent, and HCI researchers must navigate these divergent perspectives as they shape the future of AI research, regulation, and societal impact. The immediate consequences of longtermism lead to neglecting marginalized communities, exacerbating social inequalities, and undermining efforts to address ongoing harm and human rights violations. We believe that HCI's engagement with AI and its societal impact should focus on these harms and advocate for the rights and voices of the marginalized. Making design choices that contribute to this focus should become standard practice.

REFERENCES

[1] Anne Adams, Peter Lunt, and Paul Cairns. 2008. A qualititative approach to HCI research. In *Research Methods for Human-Computer Interaction*, Paul Cairns and Anna Cox (eds.). Cambridge University Press, Cambridge, UK, 138–157. Retrieved July 27, 2016 from http://www.cambridge.org/catalogue/catalogue.asp?isbn$=$9780521870122&ss$=$toc

[2] Carol J. Adams, Alice Crary, and Lori Gruen. 2023. *The Good It Promises, the Harm It Does: Critical Essays on Effective Altruism*. New York, US: Oxford University Press.

[3] Jeffrey Bardzell and Shaowen Bardzell. 2015. Humanistic HCI. *Synthesis Lectures on Human-Centered Informatics* 8, 4: 1–185. https://doi.org/10.2200/S00664ED1V01Y201508HCI031

[4] Nicholas Beckstead. 2013. On the overwhelming importance of shaping the far future. *Ph.D. thesis* Rutgers University. https://doi.org/10.7282/T35M649T

[5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21), 610–623. https://doi.org/10.1145/3442188.3445922

[6] Ruha Benjamin. 2019. *Race after technology: abolitionist tools for the new Jim code*. Polity, Medford, MA.

[7] Nick Bostrom. 2002. Existential risks: analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology* 9. Retrieved April 4, 2023 from https://ora.ox.ac.uk/objects/uuid:827452c3-fcba-41b8-86b0-407293e6617c

[8] Nick Bostrom. 2003. Astronomical waste: The opportunity cost of delayed technological development. *Utilitas* 15, 3: 308–314.

[9] Nick Bostrom. 2009. The Future of Humanity. In *New Waves in Philosophy of Technology*, Jan Kyrre Berg Olsen, Evan Selinger and Søren Riis (eds.). Palgrave Macmillan UK, London, 186–215. https://doi.org/10.1057/9780230227279_10

[10] Nick Bostrom. 2013. Why I Want to be a Posthuman When I Grow Up. In *The Transhumanist Reader*. John Wiley & Sons, Ltd, 28–53. https://doi.org/10.1002/9781118555927.ch3

[11] Victoria Braun and Virginia Clarke. 2014. Thematic analysis. In *Qualitative Research in Clinical and Health Psychology*, Poul Rohleder and Antonia Lyons (eds.). Palgrave MacMillan, Basingstoke.

[12] Therese Budniakiewicz. 1992. *Fundamentals of Story Logic: Introduction to Greimassian Semiotics*. John Benjamins Publishing.

[13] Shaan Chopra, Emma Dixon, Kausalya Ganesh, Alisha Pradhan, Mary L. Radnofsky, and Amanda Lazar. 2021. Designing for and with People with Dementia using a Human Rights-Based Approach. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI EA '21), 1–8. https://doi.org/10.1145/3411763.3443434

[14] Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. ChatGPT and the Rise of Large Language Models: The New AI-Driven Infodemic Threat in Public Health. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4352931

[15] Eva A. M. van Dis, Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L. Bockting. 2023. ChatGPT: five priorities for research. *Nature* 614, 7947: 224–226. https://doi.org/10.1038/d41586-023-00288-7

[16] Virginia Eubanks. 2017. *Automating inequality: how high-tech tools profile, police, and punish the poor*. St. Martin's Press, New York, NY.

[17] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines* 30, 4: 681–694. https://doi.org/10.1007/s11023-020-09548-1

[18] Batya Friedman, Peter H. Kahn, Alan Borning, and Alina Huldtgren. 2013. Value Sensitive Design and Information Systems. In *Early engagement and new technologies: Opening up the laboratory*, Neelke Doorn, Daan Schuurbiers, Ibo van de Poel and Michael E. Gorman (eds.). Springer Netherlands, Dordrecht, 55–95. https://doi.org/10.1007/978-94-007-7844-3_4

[19] Future of Life Institute. 2023. Pause Giant AI Experiments: An Open Letter. *Future of Life Institute*. Retrieved March 29, 2023 from https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[20] Alessandro Gabbiadini, Ognibene Dimitri, Cristina Baldissarri, and Anna Manfredi. 2023. Does ChatGPT Pose a Threat to Human Identity? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4377900

[21] Timnit Gebru, Emily M. Bender, Angelina McMillan-Major, and Margaret Mitchell. 2023. Statement from the listed authors of Stochastic Parrots on the "AI pause" letter. *DAIR*. Retrieved March 31, 2023 from https://www.dair-institute.org/blog/letter-statement-March2023

[22] Nico Grant. 2023. Google Devising Radical Search Changes to Beat Back A.I. Rivals. *The New York Times*. Retrieved from https://www.nytimes.com/2023/04/16/technology/google-search-engine-ai.html

[23] Nico Grant and Cade Metz. 2023. Google Releases Bard, Its Competitor in the Race to Create A.I. Chatbots. *The New York Times*. Retrieved from https://www.nytimes.com/2023/03/21/technology/google-bard-chatbot.html

[24] Jamie Grierson. 2023. Photographer admits prize-winning image was AI-generated. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2023/apr/17/photographer-admits-prize-winning-image-was-ai-generated

[25] Kalley Huang. 2023. Why Pope Francis Is the Star of A.I.-Generated Photos. *The New York Times*. Retrieved from https://www.nytimes.com/2023/04/08/technology/ai-photos-pope-francis.html

[26] Italian SA authority. 2023. ChatGPT: OpenAI reinstates service in Italy with enhanced transparency and rights for european users and non-users. Retrieved from https://www.garanteprivacy.it:443/home/docweb/-/docweb-display/docweb/9881490

[27] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stepha Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103: 102274. https://doi.org/10.1016/j.lindif.2023.102274

[28] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. GPT-4 Passes the Bar Exam. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4389233

[29] Reuben Kirkham. 2020. Using European Human Rights Jurisprudence for Incorporating Values into Design. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, 115–128. https://doi.org/10.1145/3357236.3395539

[30] Steve Lohr. 2023. A.I. Is Coming for Lawyers, Again. *The New York Times*. Retrieved from https://www.nytimes.com/2023/04/10/technology/ai-is-coming-for-lawyers-again.html

[31] Dieuwertje Luitse and Wiebke Denkena. 2021. The great Transformer: Examining the role of large language models in the political economy of AI. *Big Data & Society* 8, 2: 20539517211047736. https://doi.org/10.1177/20539517211047734

[32] William MacAskill. 2022. *What we owe the future*. Hachette, New York, NY.

[33] Lorna McGregor, Daragh Murray, and Vivian Ng. 2019. International Human Rights Law As A Framework For Algorithmic Accountability. *International & Comparative Law Quarterly* 68, 2: 309–343. https://doi.org/10.1017/S0020589319000046

[34] Kris McGuffie and Alex Newhouse. 2020. The Radicalization Risks of GPT-3 and Advanced Neural Language Models. https://doi.org/10.48550/ARXIV.2009.06807

[35] Cade Metz and Karen Weise. 2023. A Tech Race Begins as Microsoft Adds A.I. to Its Search Engine. *The New York Times*. Retrieved from https://www.nytimes.com/2023/02/07/technology/microsoft-ai-chatgpt-bing.html

[36] Christopher Mims. 2023. AI Tech Enables Industrial-Scale Intellectual-Property Theft, Say Critics. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/ai-chatgpt-dall-e-microsoft-rutkowski-github-artificial-intelligence-11675466857

[37] Marina Moreno. 2023. Does longtermism depend on questionable forms of aggregation? *Intergenerational Justice Review*, 1: 13–23. https://doi.org/10.24357/igjr.8.1.996

[38] Siobhan O'Connor and ChatGPT. 2023. Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse? *Nurse Education in*

Practice 66: 103537. https://doi.org/10.1016/j.nepr.2022.103537

[39] Toby Ord. 2021. *The precipice: existential risk and the future of humanity*. Bloomsbury Academic, London New York (N.Y.).

[40] Rich Pelley. 2023. 'We got bored waiting for Oasis to re-form': AIsis, the band fronted by an AI Liam Gallagher. *The Guardian*. Retrieved from https://www.theguardian.com/music/2023/apr/18/oasis-aisis-band-fronted-by-an-ai-liam-gallagher

[41] Vinodkumar Prabhakaran, Margaret Mitchell, Timnit Gebru, and Iason Gabriel. 2022. A Human Rights-Based Approach to Responsible AI. https://doi.org/10.48550/arXiv.2210.02667

[42] Jürgen Rudolph, Samson Tan, and Shannon Tan. 2023. ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching* 6, 1. https://doi.org/10.37074/jalt.2023.6.1.9

[43] Katharine Sanderson. 2023. GPT-4 is here: what scientists think. *Nature* 615, 7954: 773–773. https://doi.org/10.1038/d41586-023-00816-5

[44] Adam Satariano. 2023. ChatGPT Is Banned in Italy Over Privacy Concerns. *The New York Times*. Retrieved from https://www.nytimes.com/2023/03/31/technology/chatgpt-italy-ban.html

[45] Chirag Shah and Emily M. Bender. 2022. Situating Search. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval* (CHIIR '22), 221–232. https://doi.org/10.1145/3498366.3505816

[46] Filipo Sharevski, Jennifer Vander Loop, Peter Jachim, Amy Devine, and Emma Pieroni. 2023. Talking Abortion (Mis)information with ChatGPT on TikTok. https://doi.org/10.48550/ARXIV.2303.13524

[47] Ahmed Tlili, Boulus Shehata, Michael Agyemang Adarkwah, Aras Bozkurt, Daniel T. Hickey, Ronghuai Huang, and Brighter Agyemang. 2023. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments* 10, 1: 15. https://doi.org/10.1186/s40561-023-00237-x

[48] Adrienne Williams, Milagros Miceli, and Timnit Gebru. 2022. The Exploited Labor Behind Artificial Intelligence. *Noema*. Retrieved from https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence

[49] Eliezer Yudkowsky. 2023. Pausing AI Developments Isn't Enough. We Need to Shut it All Down. *Time*. Retrieved April 11, 2023 from https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/

[50] Shoshana Zuboff. 2019. *The age of surveillance capitalism: the fight for a human future at the new frontier of power*. PublicAffairs, New York.