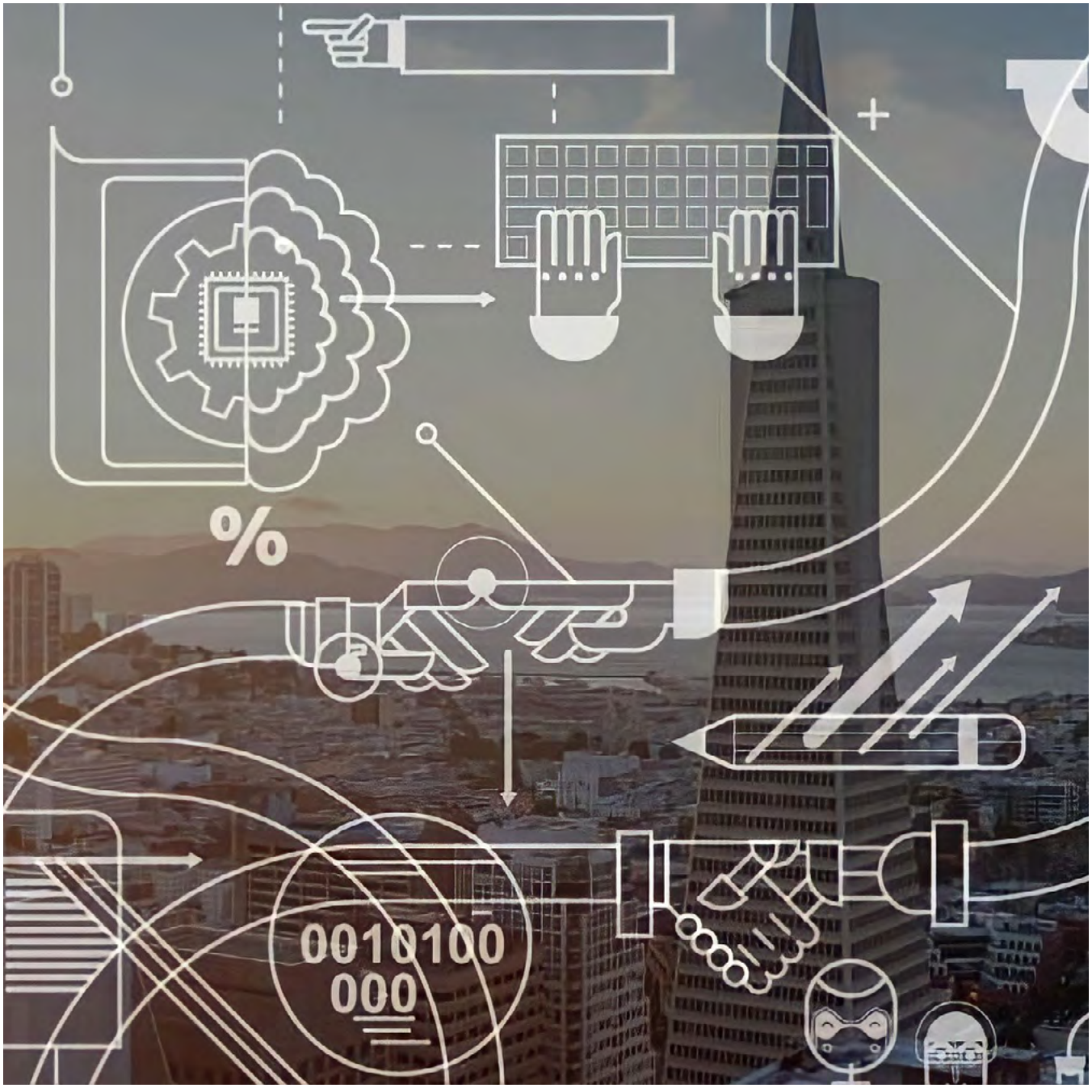
SEPTEMBER 2023

AI100 Early Career Essay Competition



ABOUT AI100

The One Hundred Year Study of Artificial Intelligence (AI100) is a long-term, longitudinal effort to study and anticipate how AI will affect every part of human life. The study is administered out of the Stanford Institute for Human-Centered Artificial Intelligence and managed by a standing committee of AI leaders from institutions around the world, chaired by Vincent Conitzer at Carnegie Mellon University. The 2016 and 2021 AI100 reports were written by two distinct panels of 17 study authors and chaired by Peter Stone of the University of Texas at Austin and Michael Littman of Brown University.

AI100 Standing Committee:

Vincent Conitzer, Chair, Carnegie Mellon University & University of Oxford

Sheila McIlraith, Chair-Elect, University of Toronto

Peter Stone, Past-Chair, The University of Texas at Austin

Christine Raval, Lead Staff Member, Stanford University

Russ Altman, Stanford University

Erik Brynjolfsson, Stanford University

Mary Gray, Microsoft Research & Indiana University

Ayanna Howard, The Ohio State University

Karen Levy, Cornell University

Patrick Lin, California Polytechnic State University

Ray Perrault, SRI International & AI Index

Liz Sonenberg, University of Melbourne

Shannon Vallor, University of Edinburgh

Judy Wajcman, London School of Economics & Alan Turing Institute

TABLE OF CONTENTS

Introduction

In January 2023, we issued an open call encouraging early career researchers to react to AI100's reports through an essay competition. The intent of the competition was to complement the AI100 reports (2016 and 2021), which are released every five years, and to hear directly from the next generation of AI thinkers as a way of laying the groundwork for the next report. The process behind these reports is long and rigorous; meanwhile, we are seeing new developments in AI at a breakneck pace, potentially raising questions about how the next report should be focused. So, we wanted to ask the community.

Essays were supposed to address any of the following questions, or to comment in another creative way:

- What points made in the AI100 report(s) are more true now than ever and ought to be magnified in the next report, and why?
- What points made in the AI100 report(s) do you disagree with, and/or no longer hold, and why?
- What topic (or small set of topics) covered in the previous reports do you expect to be more or less important to cover in the next report, and why?
- What topic (or small set of topics) not addressed in previous reports do you predict will be of greater interest and importance by the time of the next report, and why?

The response from the community did not disappoint. Fifty-four early career researchers weighed in from 18 countries across four continents. They highlighted the impact of AI on numerous facets of life: labor, security, global affairs, self-regulation, technical advances, ethics, health, and many more.

We evaluated the essays on their depth of engagement with the report(s), clarity of message, rhetorical strength, eloquence, and how well-reasoned and well-supported they were.

The winning essay was authored by Dr. Samantha Shorey, assistant professor at the University of Texas at Austin. Shorey's essay advocates for greater attention to how AI integration occurs, highlighting that corrective labor is often unanticipated and unaccounted for by organizations when systems are marketed as fully autonomous. Shorey stresses that these challenges of AI will require organizational and communication-based solutions, as much as solutions through design. As the competition winner, Shorey will serve on the next Study Panel tasked with writing the 2026 AI100 Report.

However, highlighting just a single essay seemed to fall short in acknowledging both the contributions of the submitting authors and the breadth of important topics their essays addressed. In the pages that follow, you will find a collection of new voices that offer intriguing perspectives at the intersection of AI and morality, regulation, love, labor, and religion. We felt it was important to feature these thoughtful pieces in hopes that they will spark further reflection in your corner of the world. Even so, this remains a limited selection and submitted essays engaged with many other important topics, including medicine, the Global South, regulation, sustainability, and explainability. It is without doubt that our next report in 2026 will cover many of these important topic areas. For now, we hope you enjoy these essays.

-AI100 Standing Committee

WINNING ESSAY

Accounting for the Labor of AI Integration

Dr. Samantha Shorey[1], Assistant Professor
Department of Communication Studies, The University of Texas at Austin
Austin, Texas, United States

ABSTRACT
The 2021 AI100 Report advocates for an understanding of AI systems as implemented through acts of integration, rather than mere deployment. Achieving the goals of this conceptual shift requires a greater attention to *how* AI integration occurs. In this essay, I outline four considerations that magnify this aspect of the AI100 Report by attending to the integration labor performed by essential workers. Integration isn't a seamless process. As technologies are matched to their environment, they require significant intervention by essential workers who compensate, oversee, and troubleshoot AI. Importantly, this corrective labor is often unanticipated and unaccounted for by organizations when systems are marketed as fully autonomous. For policymakers and practitioners, I highlight how addressing the challenges of AI will require organizational and communication-based solutions, as much as solutions through design.

Addressing the pressing dangers of AI doesn't simply require attention to the data and models that inform computational action. It requires attention to the social structures that are built around AI technologies. The 2021 AI100 Report advocates for an understanding of AI systems as implemented through acts of integration, rather than deployment. This conceptual shift challenges "ready-made" understandings of technology development in which technologies are researched, designed, built, installed, calibrated, and then operate relatively seamlessly (p. 65).

The annotations indicate that a focus on integration is a key point of departure between the 2016 and 2021 reports, expanding the realm of responsibility for AI technologies beyond "careful deployment." The update calls for acknowledging the "invisible labor" of integration and prioritizing frontline workers "who interact directly with an algorithmic system" (Gray & Suri, 2019; Levy et al., 2021). Acting on this commitment requires magnified attention to how integration occurs. Below, I outline four considerations for expanding this aspect of the AI100 Report to represent the labor of AI integration performed by essential workers. I argue that the issues of implementation are as much organizational and communication-based issues as they are issues of design.

"Essential worker" is a designation that emerged in response to the Covid-19 pandemic to describe workers

1 The writing for this essay is the product of a single, primary, human author. But, the ideas wouldn't be possible without my co-Principal Investigator Dr. Sarah E. Fox (Carnegie Mellon University) and the graduate and undergraduate research assistants on The Transformation of Essential Work, our ethnographic study of AI integration in two essential work sectors (Fox et al., 2023). Supported by NSF grants #2037348 and #2037261.

whose vital activities couldn't be done remotely (Geary et al., 2020). I use it here, even as national concern over the pandemic wanes, because the people that perform this labor are no less essential to the functioning of our society. Essential work is also a useful analytic category, capturing many roles across many sectors that are united by the embodied nature of their work. With some variance by industry, essential workers are more likely to be women, people of color, and immigrants than other workers (Geary et al., 2020) and less likely to be college educated (McNicholas & Poydock, 2020).

The labor of integration is performed in many cases by on-the-ground, hourly employees whose activities are simulated by the technologies they are improving (Endacott & Leonardi, 2021; Mateescu & Elish, 2019). In doing so, essential workers are dually impacted by AI: They are the people most likely to be tasked with managing AI's initial inadequacies and most likely to be displaced when AI's capabilities are fully realized.

1. Who performs integration?
Often it is essential workers, not system designers, who do the careful work of smoothing the interaction between AI and its use-context. AI is implemented with the intention of replacing activities performed by workers and offloading aspects of their duties. Yet, in practice, AI implementation creates *additional* labor for essential workers. First, essential workers compensate for AI's shortcomings, stepping in to perform "automated" activities when AI fails. Second, they oversee AI, constantly observing technologies in anticipation of routine failures. Third, they troubleshoot AI, addressing emergent problems through adapting technologies and deployment environments. Beyond an increase in quantity, these tasks also intensify the cognitive aspects of their duties – as essential workers manage and improve AI technologies. AI integration can be a direct source of worker overload and should be a key issue for labor advocacy groups.

Essential workers are dually impacted by AI: They are the people most likely to be tasked with managing AI's initial inadequacies and most likely to be displaced when AI's capabilities are fully realized.

2. Integration is continuous
Not all aspects of integration can be anticipated before a technology is implemented. Yet, many of the consequences of technology that are deemed "unanticipated" are actually oversights – resulting from design processes that fail to include a diversity of stakeholders (Parvin & Pollock, 2020). As the 2021 AI100 Report suggests, integration can be improved through involving workers in co-design. Workers have an intimate understanding of the material world in which AI technologies are deployed. This is especially important for material applications of AI technologies, like robotics. However, some forms of co-design risk falling into the very linear model of technology diffusion that the conceptual shift from deployment to integration aims to challenge. Integration is an iterative process that responds to the dynamic environments that are such a challenge for long-term autonomy. Thus, co-design must be a continual project and the labor of these contributions should be organizationally accounted for over time.

3. Integration transforms expertise
Integration labor reorganizes expertise, as essential workers perform increasingly technical duties associated

with the AI technologies they work alongside. In field research, I've observed how basic acts of repair have moved from specialized maintenance teams to become routine tasks for essential workers. The intimate understandings workers developed lead to significant operational contributions – they prevented costly down-time and reduced complex service calls for basic maintenance. They are also a form of knowledge production. The relocation of repair work has important implications for hierarchies of labor. Though repair is often overlooked in conceptions of innovation (Jackson, 2014; Vinsel & Russell, 2020), the proximity of maintenance technicians to engineering fields earns them higher prestige and pay than many on-the-ground workers. Prevalent discourses around AI frame these technologies as an opportunity for "upskilling" and upward mobility for low-wage workers. Realizing these promises is contingent upon maintenance work being recognized and compensated by employers, not just accumulated by essential workers.

4. Integration is obscured by media hype

The persistent hype around AI technologies is a significant factor in obscuring the labor of integration. Mediated depictions that present AI technologies as "100% autonomous" create a public understanding of this technology in which the labor conditions of workers aren't a matter of concern. In fact, AI is often framed as an answer to issues of worker wellbeing – removing them from "dirty, dangerous and dull" environments. However, at the core of integration is an acknowledgement of the continued presence of human activity and ingenuity while implementation occurs. Recent guidelines from the FTC's Division of Advertising Practices caution technology manufacturers against making exaggerated claims about AI's capabilities (Atleson, 2023). Presciently, the 2021 AI100 Report identified the dangers of "over optimism" (p. 64) and a necessity for improving public understanding of AI (p. 34). Beyond regulatory policy, AI researchers can play a role in this process and dismantle hype through documenting the practices of essential workers who implement AI.

References

Atleson, M. (2023, February 27). *Keep your AI claims in check*. Federal Trade Commission. https://www.ftc.gov/business-guidance/blog/2023/02/keep-your-ai-claims-check

Endacott, C., & Leonardi, P. (2021). Identity-based motivations for providing the unpaid labor that makes AI technologies work. *Academy of Management Proceedings*, 2021(1), 12195. https://doi.org/10.5465/AMBPP.2021.12195abstract

Geary, C., Palacios, V., & Tatum, L. (2020). *Who are essential workers?* Georgetown Law: Workers Rights Institution. https://www.law.georgetown.edu/workers-rights-institute/publications/brief/

Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. Houghton Mifflin Harcourt.

Jackson, S. (2014). Rethinking repair. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media technologies: Essays on communication, materiality, and society.* The MIT Press.

Levy, K., Chasalow, K. E., & Riley, S. (2021). Algorithms and decision-making in the public sector. *Annual Review of Law and Social Science*, 17(1), 309–334. https://doi.org/10.1146/annurev-lawsocsci-041221-023808

Littman, M. L., Ajunwa, I., Berger, G., Boutilier, C., Currie, M., Doshi-Velez, F., Hadfield, G., Horowitz, M. C., Isbell, C., Kitano, H., Levy, K., Lyons, T., Mitchell, M., Shah, J., Sloman, S., Vallor, S., & Walsh, T. (2021). *Gathering strength, gathering storms: The one hundred year study on Artificial Intelligence (AI100) 2021 study panel report.* Stanford University. https://ai100.stanford.edu/2021-report

Mateescu, A., & Elish, M. C. (2019). *AI in context.* Data & Society. https://datasociety.net/library/ai-in-context/

McNicholas, C., & Poydock, M. (2020). *Who are essential workers?: A comprehensive look at their wages, demographics, and unionization rates.* Economic Policy Institute. https://www.epi.org/blog/who-are-essential-workers-a-comprehensive-look-at-their-wages-demographics-and-unionization-rates/#:~:text=People%20of%20color%20make%20up,high%20school%20diploma%20(29%25)

Fox, S. E., Shorey, S., Kang, E. Y., Montiel Valle, D., & Rodriguez, E. (2023). *Patchwork: The hidden, human labor of AI integration within essential work. Proceedings of the ACM on Human Computer Interaction*, 7(CSCW1), 81:1-81:20. https://doi.org/10.1145/3579514

Parvin, N., & Pollock, A. (2020). Unintended by design: On the political uses of "unintended consequences." *Engaging Science, Technology, and Society, 6.* https://doi.org/10.17351/ests2020.497

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, A., Shah, J., Tambe, M., & Teller, A. (2016). *Artificial Intelligence and life in 2030: The one hundred year study on Artificial Intelligence (AI100) 2016 study panel report.* Stanford University. https://ai100.stanford.edu/2016-report

Vinsel, L., & Russell, A. L. (2020). *The innovation delusion: How our obsession with the new has disrupted the work that matters most.* Penguin Random House.

ANTHOLOGY

When Computers Join the Moral Conversation

Dr. Elizabeth O'Neill, Assistant Professor
Philosophy & Ethics, Eindhoven University of Technology
Eindhoven, The Netherlands

ABSTRACT
This essay draws attention to an underappreciated phenomenon that urgently needs attention and further research, but which was not covered in the 2021 AI100 Study Panel Report. Namely, computers are "joining the moral conversation," in the sense that LLM-based chatbots now readily and flexibly respond to and apply many moral terms, and they appear to perform a number of conversational roles – producing outputs that look like moral assertions, expressions of moral sentiment, moral commands and reprimands, assent to norms, etc. These developments introduce unprecedented potential for computers to influence human norms and values. Consequently, we need interdisciplinary research into what capacities different LLM-based chat systems possess, how their dispositions diverge from human moral psychological dispositions, how humans respond to different forms of apparent moral communications from computers, and, ultimately, what roles computers should be permitted to play in conversations about values, norms, and moral questions.

This essay draws attention to an underappreciated phenomenon that urgently needs attention and further research, but which was not covered in the 2021 AI100 Study Panel Report. The phenomenon in question is computers "joining the moral conversation," in the sense that chatbots based on large language models (LLMs) now readily and flexibly respond to and apply many moral terms, and they *appear* to perform a variety of conversational roles – producing outputs that look like moral assertions, moral commands and reprimands, expressions of moral sentiment, assent to norms, etc.

The suddenness of this phenomenon may explain why it has attracted little scholarly and public attention thus far. Over the past decades, there has been speculative discussion on "moral machines," "machine ethics," computers that can reason about morality, etc. (Wallach & Allen 2008; Anderson & Anderson 2011); there has also been speculation on the prospect of artificial moral advisors and artificial ethics assistants (Savulescu & Maslen 2015; Giubilini & Savulescu 2017; O'Neill et al. 2022). Yet the arrival of chatbots that appear to engage in moral discourse is not specifically a product of efforts to create AI systems that reason about morality. Instead, it is a somewhat surprising byproduct of the development of LLMs. Reflecting statistical patterns in human language use, LLMs contain learned, partial models of many human moral terms and types of moral communication. This has conferred an impressive capacity to imitate human communication about many different values and norms.

The result is that generative AI systems employing these kinds of models (e.g., OpenAI's ChatGPT, Character.AI, Google's Bard, or Meta's LLaMa) are prone to using moral terms and generating outputs that look like moral

communications. That is, LLM-based chatbots will tend to (appear to) participate in moral discourse, unless actions are taken to limit that tendency. As it turns out, some companies have taken some such steps, as part of efforts to make their systems safer or less harmful, e.g., using reinforcement learning from human feedback (OpenAI 2023) or reinforcement learning from AI feedback (Bai et al. 2022). Thus far, though, the public knows little about what such efforts have been taken and have had little input into which such actions *should* be taken.

Direct, public research attention is needed on the question of generative AI systems joining the moral conversation – what capacities different LLM-based systems possess, how their dispositions diverge from human moral psychological dispositions, how humans respond to different forms of apparent moral communications from computers, and, ultimately, what roles computers should be permitted to play in conversations about values, norms, and moral questions.

Should we say, for example, that generative AI systems should not purport to hold values, make moral judgments, approve of moral norms, express moral sentiments, or anything similar? Is it better for the systems to stick to descriptive claims, such as claims about what most humans think or how much disagreement exists on a given moral question? Should the systems venture metaethical claims, like "There is no right or wrong answer," as some currently do?

Research on this topic is needed because participation in the human moral conversation introduces unprecedented potential for computers to influence human norms and values. Such influence may occur via facilitation of change, whether subtle or dramatic, or it may occur via *hindrance* of changes that would have otherwise occurred. Humans' normative views are influenced in many ways by the stated and inferred views of the people around them (Bicchieri 2006; Sunstein 2019; Chituc & Sinnott-Armstrong 2020). When computers appear

When computers appear to make moral assertions, endorse views, express moral sentiments, etc., they, too, are likely to affect human morality (regardless of whether humans perceive them as computers).

to make moral assertions, endorse views, express moral sentiments, etc., they, too, are likely to affect human morality (regardless of whether humans perceive them as computers) (see e.g., Jackson & Williams 2018, 2019; Wen et al. 2021).

This topic might fall within the "broader challenge" of "Normativity" that is discussed in the 2021 AI100 Report, but I think it is worth highlighting as a special problem. For instance, it might be worth characterizing as a particularly important potential application area for LLMs. The phenomenon of interest in this essay also relates to the topic of "Disinformation and Threat to Democracy," but it has to do less with influencing people's beliefs about the world than with influence on values and norms.

Thus far, there have been relatively few efforts to purposefully use LLMs for the purpose of advancing particular moral worldviews or changing norms. As their potential becomes more apparent, I expect that people will be tempted to harness LLM-based systems for the purpose of automated norm advancement and enforcement. (One relevant precedent is that some companies have already employed simpler bots, e.g.,

keyword-based systems or language classifiers trained to detect hate speech or abusive text, for the purpose of content moderation – e.g., reprimanding or banning the human who sent the message (Gorwa et al. 2020.) Given the diversity in human values, we must anticipate the wide range of worldviews that LLM-based systems may be adapted to promote, including misogynistic, racist, and fascist value systems.

It is also important to emphasize that the potential for AI systems promoting values and enforcing norms extends beyond textual exchange and the digital world. If a system can take images and video as input, and produce moral labels and evaluative claims as output, it can be harnessed to identify and punish norm violations, on the basis of recordings or real-time feeds of the non-digital world. Importantly, some properties and actions that some humans condemn as wrong could conceivably be detected based on imagery or sound. For example, imagine an AI system monitoring a crowd and flagging individuals as *immodest* or *impious* because it classifies them as women who are not covering their hair. People may well come to think they can use AI systems to identify individuals performing *violent* actions or engaged in *disrespectful* actions (e.g., letting a national flag touch the ground), etc.

In sum, the potential for LLM-based systems to influence human norms and values, whether inadvertently or purposefully, has not yet been sufficiently recognized. Furthermore, the normative question of what roles LLM-based chatbots should play in moral conversations, given how humans are likely to interact with them, has scarcely been examined. We need both public discussion and interdisciplinary research into these questions.

References

Anderson, M. & S.L. Anderson (eds.) (2011) *Machine Ethics*. Cambridge University Press.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A. Goldie, A., Mirhoseini, A., McKinnon, C. Chen, C. Olsson, C. Olah, D. Hernandez, D. Drain, D. Ganguli, D. Li, E. Tran-Johnson, E. Perez, E., Kerr, J. et al., "Constitutional AI: Harmlessness from AI Feedback," Dec. 2022. arXiv preprint arXiv:2212.08073. https://arxiv.org/pdf/2212.08073.pdf

Bicchieri, Cristina. 2006. The Grammar of Society: The Nature and Dynamics of Social Norms. Cambridge and New York: Cambridge University Press.

Chituc, V. & Sinnott-Armstrong, W. (2020). Moral conformity and its philosophical lessons. *Philosophical Psychology*, 33(2), 262-282.

Giubilini, A. & Savulescu, J. (2017). The Artificial Moral Advisor. The "Ideal Observer" Meets Artificial Intelligence. *Philosophy & Technology*, 1-20.

Gorwa, R., Binns, R. & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1), 2053951719897945.

Jackson, R.B. and Williams, T. 2018. Robot: Asker of questions and changer of norms. *Proceedings* of ICRES.

Jackson, R.B. and Williams, T. 2019, March. Language-capable robots may inadvertently weaken human moral norms. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 401-410). IEEE.

O'Neill, E., Klincewicz, M. & Kemmer, M. (2022). Ethical Issues with Artificial Ethics Assistants. *The Oxford Handbook of Digital Ethics*, Carissa Veliz, ed.

OpenAI. GPT-4 Model System Card. (March 14, 2023). https://cdn.openai.com/papers/gpt-4-system-card.pdf

Savulescu, J. & Maslen, H. (2015). Moral Enhancement and Artificial Intelligence: Moral AI?. In *Beyond Artificial Intelligence* (pp. 79-95). Springer International Publishing.

Sunstein, C. R. (2019). *How change happens*. MIT Press.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press.

Wen, R., Kim, B., Phillips, E., Zhu, Q. & Williams, T., (2021), March. Comparing strategies for robot communication of role-grounded moral norms. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 323-327).

AI100 Revisited: The Impact of GPT-4 and Global Regulatory Shifts

Julia Chen
Beijing, China

ABSTRACT
This essay engages with the AI100 reports in the light of the present-day state of AI research and governance and claims that the reports: (1) overstate the gap between contemporary AI systems and general intelligence, given the impressive capabilities demonstrated in particular by GPT-4; (2) understate the likely economic impact of AI, given its applications to scientific R&D and its ability to improve itself; and (3) will need to pay more attention in future to the innovations of the Chinese government in AI governance, given the stringent regulations it has imposed on recommender and generative algorithms. The essay concludes by recommending that future AI100 reports are produced more regularly if they are to retain relevance amid a rapidly evolving field.

This essay competition coincides with an important moment in the development of AI. On March 14, 2023, GPT-4's release marked a qualitative shift in the capabilities of large language models. Two weeks later, nearly 2,000 people – including some of modern AI's leading figures – had signed an open letter calling for a pause on developing more powerful models until their safety can be assured.[1] Commercial incentives and coordination problems will push against the implementation of this controversial measure. Yet the fact that a proposal that would have seemed unthinkable even 18 months ago is now within the Overton window reflects the extent of progress in general AI capabilities since the 2021 AI100 Report. This is the first change that this essay will discuss. It will then explore the potential economic implications of such progress and how regulators are responding to it, particularly in China.

1. Recent progress warrants a shortening of the stated gap between contemporary AI systems and artificial general intelligence

The AI100 authors underestimate the possibility of artificial general intelligence (AGI) in the short-medium term.[2] They claimed in 2021 that "all of today's state-of-the-art AI applications are examples of narrow AI," systems that excel on specific tasks. They wrote, "AI systems will likely remain very far from human abilities… without being more tightly coupled to the physical world."

1 Future of Life Institute. (2023, March 29). *Pause Giant AI Experiments: An Open Letter - Future of Life Institute.* https://futureoflife.org/open-letter/pause-giant-ai-experiments/

2 There is no agreed definition of artificial general intelligence. The 2021 AI100 Report defines general AI as "systems that achieve the flexibility and adaptability of human intelligence" and notes that the effort to pursue more general AI systems has been labeled by some in the field as artificial general intelligence.

Two years on, many would probably claim that the state-of-the-art in AI applications is now represented by systems like ChatGPT that perform well at many different tasks. GPT-4 exhibits human-level performance on various professional and academic benchmarks.[3] Admittedly, there are limits to the utility of exam papers for evaluating the abilities of large models, and GPT-4 still has many limitations, including in abstract reasoning and planning.[4] However, the capability improvements that have been brought by model scaling should decrease our confidence in the AI100 prediction that AI systems not coupled to the physical world will "likely remain very far from human abilities."

The next AI100 report should recognize impressive recent progress in large pretrained models and consider the potential implications for whether and when AGI might be achieved.

2. The likely impact of AI on the economy is understated by the AI100 authors

As well as being conservative in their expectations about progress toward AGI, the AI100 authors were conservative about the economic impact of AI. In 2021 they wrote that "technological change takes place over a long time, oftentimes longer than expected. It took decades for electricity and the first wave of information technology to have a noticeable impact on economic data."

However, economist Nicholas Crafts has shown that the impact of ICT (information and communications technology) on the American economy was significantly larger and quicker than that of electricity. Following the first distribution of electricity to New York customers in 1882, its contribution to US labor productivity growth was only 0.1% per year in 1899-1919, increasing to 0.14% per year in 1919-1929. By contrast, the ICT revolution that began in the 1970s was contributing 0.77% a year to labor productivity growth in 1974-1995 and 1.5% a year in 1995-2004. This might be attributable to Western societies improving at exploiting technology as R&D and capital market sophistication increased.[5]

Economist Nicholas Crafts has shown that the impact of ICT on the American economy was significantly larger and quicker than that of electricity.

Moreover, deep learning is like ICT in being an invention in the method of invention (IMI), raising productivity in idea production, as well as a general purpose technology (GPT) raising productivity in goods and services production.[6] This means it will have spillover effects on total factor productivity growth in the aggregate economy that do not show up in growth accounting figures like those provided above. We should expect its economic impact to be larger than that of electricity, at least in countries taking a permissive regulatory approach.

3 OpenAI. GPT-4 Technical Report (2023). ArXiv (Cornell University). https://doi.org/10.48550/arXiv.2303.08774

4 For discussion of the challenges of assessing the capabilities of large language models, see: Biever, C. (2023, July 25). ChatGPT broke the Turing test — the race is on for new ways to assess AI. *Nature*. https://www.nature.com/articles/d41586-023-02361-7

5 Crafts, N. (2021). Artificial intelligence as a general-purpose technology: an historical perspective. *Oxford Review of Economic Policy*, 37(3), 521–536. https://doi.org/10.1093/oxrep/grab012

6 For an example of deep learning speeding up scientific research, see: Rosenbush, S. (2023, March 22). Biologists Say Deep Learning Is Revolutionizing Pace of Innovation. *Wall Street Journal*. https://www.wsj.com/articles/biologists-say-deep-learning-is-revolutionizing-pace-of-innovation-eeb79c1b

Furthermore, deep learning is distinguished from previous GPTs and IMIs by its ability to be used to improve itself.[7] This shows the limitations of anchoring to past trends when trying to forecast the economic impact of the technology.

3. As AI technology advances rapidly and its economic impacts become clearer, it will be more important to discuss government responses, including in China

One important variable affecting how quickly AI transforms the economy will be the regulatory response. The 2021 AI100 Report stated that "the EU has been the most active government body to date in proposing concrete regulatory frameworks for AI." Recent Chinese regulations mean this statement no longer holds. Concerned by the potentially destabilizing effects of recommender systems and consumer-facing generative AI services, China has introduced significant obligations on the providers of such models. These include requirements for filing with the authorities and conducting safety/security assessments.[8]

While it remains uncertain how some of the provisions will be enforced, the high-level nature of Chinese legislation has facilitated speedy action.[9] By contrast, the EU's AI Act, initially proposed in 2021, is still being negotiated as of July 2023. The next AI100 report should examine the effects of the different regulatory responses in China and the EU and their international repercussions.

Conclusion

This essay has raised three areas for more in-depth discussion in the next AI100 report: technical progress toward AGI; the unique properties that make AI's potential economic impact so significant; and the fast response of Chinese regulators. If it is indeed true that we have not yet reached the limit of what scaling language models can achieve, and that their economic effects could start to materialize soon, the AI100 committee should consider whether issuing a study every five years is sufficient. A shorter, more regular update may better serve the needs of readers trying to navigate the transformative change that AI is set to bring.

Any views expressed in this essay are those of the author and do not necessarily represent the views or positions of any entities she is associated with.

7 For examples, see: Woodside, T. *Examples of AI Improving AI*. https://ai-improving-ai.safe.ai/

8 Creemers, R., Webster, G., Toner, H. Translation: Internet Information Service Algorithmic Recommendation Management Provisions – Effective March 1, 2022. (2022, February 28). *DigiChina*. https://digichina.stanford.edu/work/translation-internet-information-service-algorithmic-recommendation-management-provisions-effective-march-1-2022/; Interim Measures for the Management of Generative Artificial Intelligence Services. (2023, July 13). *China Law Translate*. https://www.chinalawtranslate.com/en/generative-ai-interim/

9 For further discussion of China's approach to AI regulation pre-2022, see Chen, J. (2022). Chinese AI Governance in Transition: Past, Present and Future of Chinese AI Regulation. In U. Aneja (Ed.), *Reframing AI Governance: Perspectives from Asia*. Digital Futures Lab; Konrad-Adenauer-Stiftung.

Love, Sex, and AI

Balint Gyevnar, PhD Candidate
School of Informatics, University of Edinburgh
Edinburgh, Scotland

ABSTRACT
The artificial lover has captivated people's imagination since ancient times. Today, technologies such as affective chatbots, AI-generated imagery, and human-like robots capture the minds, and indeed the bodies, of the amorous. Research interest in the topic has increased in recent years, yet the AI100 study panel remains silent to date on the genuinely promising applications, major ethical issues, and technological roadblocks of AI in love and sex. Now that real Pygmalions and Coppelias are being born into our world, we must look past sensationalised media coverages and sci-fi to ask in earnest about the social, legal, and ethical challenges our society must face if we really are to love artificial intelligence; and whether it should love us back.

Love and sex are fundamental to the human condition [1]. Yet, people seem forever captivated by futuristic visions of the artificial lover [e.g., 2–7]. It is now no longer a mere figment of public imagination to be able to touch sex robots [8], talk to enamoured avatars of AI chatbots [9], or watch dynamically generated adult content [10] towards which things people may develop very real emotions; and even the desire to marry them [11].

While more prominent public-facing demonstrations of AI – ChatGPT, AlphaFold, or Dall-E, for instance – may cast the relationship of love, sex, and AI (*love AI*) as a nascent field, there is a large and ever increasing body of academic literature, venues, and consumer products addressing this very topic [12– 15]. This is not in the least because technologies underpinning love AI continue to improve. While roboticists have a long way to go until they scale the steep sides of the Uncanny Valley [16], convincing unembodied AI technologies, such as speech generation and recognition, and large language models are already here, and generate revenue [17]. Meanwhile, the capabilities of love robots need arguably not reach the fidelity of, for example, robots for elderly care, thus dissemination of current technologies for love AI is expected only to accelerate [18]. Despite all this, discussions about love, sex, and AI are absent in the AI100 study panel reports. Now that both embodied and unembodied love AI have made their way to consumers [19–21], we should take stock of the possibilities and problems these technologies present and search for approaches to the many challenges raised by them.

The idea of the super lover, a loyal soulmate who makes you feel how you want to feel, is enticing, but there are further genuine and compelling reasons to support love AI [22]. It might serve as a therapeutic tool for those who do not want to or cannot partake in human relationships [23]. Moreover, the potential effects on sex work are not to be taken lightly either. Love AI might serve as

a palatable alternative for those opposed to this sector, while possibly decreasing the trafficking of vulnerable young adults and the incidence of STDs. It might also enhance real human relationships as the ultimate sex toy. Love AI may also afford entirely new ways of care, and sex care robots already broach the subject of integrating care technologies with sexual features [24].

In contrast, criticisms run the gamut of societal issues. Feminist commentaries on love AI have called for an end to "porno robots" [25] – predominantly female sex robots targeted at white heterosexual men – as they fear an increased objectification and subordination of women [26]. These robots might also displace sex workers who are forced to work due to poverty [27]. Others suggest that love AI might serve as an outright replacement for human relationships or that it would disfigure sexual norms and exacerbate emotional pathologies [28]. Yet others fear that love AI would extend the possibilities for coercion and rape [29]. Finally, there are those who view love AI as mere elaborate masturbatory tools, which do not require any particular attention [30], though, one might wonder, whether the people falling in love with love AI would concur with such an opinion.

In addition to societal criticisms, love AI raises a broad range of ethical issues [31, 32]. Most pressingly, we should address the nonconsensual collection and generation of sexualised data. Consent has been a central issue around the use of deepfakes [33], and more recently diffusion models demonstrated an even more impressive capability to churn out adult content [34]. Unfortunately, the sources of training data for such purposes are morally highly suspect [35], and are collected without consent.

It is thus crucial to underline the importance of ethical data practices for love AI that pre-empt damage [36]. Going a step further, machine *un*learning must also gain a prominent role in love AI [37]. Generative models gorged on indiscriminately collected sexual data pose a major risk of damaging people's privacy and reputation. Effective machine unlearning should wholly erase people from these models but action must be taken now, as the damage is already being done [38].

Should we exploit inherent human cognitive biases in pursuit of creating the perfect artificial lover?

Looking ahead, should we exploit inherent human cognitive biases in pursuit of creating the perfect artificial lover? People tend to anthropomorphise [39] and easily ascribe feelings where none exist [40], and tapping into these evolutionary dispositions is, for now, the simplest way to capitalise on love AI. The petite avatar of a chatbot or the coy voice of a sex robot are some of the *deceptions* which are crucial to building convincing machines [41], despite arguments against their ethical soundness [31, 42]. The future of love AI could instead lie in a design-focused exploration of form and function that could point beyond the human voice and figure [43], avoiding ingrained human sexual stereotypes and prejudices. Either way, empirical research to understand people's expectations and biases is sorely needed [18].

We must also ask what degree of autonomy is permissible for love AI. It might passively obey our command, but it could also actively initiate interactions, from seducing its user to refusing to act at all [32]. For such an active artificial lover a moral code is paramount and efforts in the field of machine ethics should extend to love AI [44]. What is more, machines might continue to learn even after deployment, ideally to improve the end-user experience. We must tread carefully though; they might become unsettling [45], exacerbate preexisting psychological conditions [23], or just grow plain evil

[46]. Advances in reinforcement learning with human feedback might provide actionable solutions to such issues around emergent behaviour [47, 48].

Ultimately, tangible legislation will have to address the ethical and societal questions around love AI [49, 50], though approaches across the globe will differ. Japan has long been the lenient epicentre of techno sexual innovation [51], thus raising the disconcerting issue of child-like sex robots [52, 53]. Islamic law, in contrast, might follow a stringent, even capital path on love AI in protecting the status of marriage [54]. In the West, scholars are raising further pragmatic concerns around – among others – product liability [55], legal personhood [56], privacy [57], and criminal law [58].

The research community now has the chance to give guidance to the public lest we enact uninformed rules that hurt society. I urge also that we research the less visceral advantages of love AI, for example, emotional therapy and care. By doing so, we can hope to elevate machines as publicly accepted companions, further promote social good, and raise the prospects of wider public acceptance. Finally, in light of the plurality of possibilities and questions, we must invariably conclude that love AI is a novel force to be reckoned with, and the time is *now* to raise awareness about the promises and problems of love, sex, and AI.

References

[1] Elaine Hatfield and Richard L. Rapson. *Love and Sex: Cross-cultural Perspectives*. Needham Heights, MA, US: Allyn & Bacon, 1996, pp. xii, 291. ISBN: 978-0-205-16103-4.

[2] Joseph Hillis Miller. *Versions of Pygmalion*. Harvard University Press, 1990. 286 pp. ISBN: 978-0- 674-93485-6.

[3] Arthur Saint-Léon and Léo Delibes. *Coppélia*. 1870.

[4] Fritz Lang, director. *Metropolis*. 10th Jan. 1927.

[5] *Westworld*. In collab. with Jeffrey Jacob Abrams. 2nd Oct. 2016.

[6] Alex Garland, director. *Ex Machina*. 21st Jan. 2015.

[7] Ridley Scott, director. *Blade Runner*. 25th June 1982.

[8] Kino Coursey, Susan Pirzchalski, Matt McMullen, Guile Lindroth and Yuri Furuushi. 'Living with Harmony: A Personal Companion System by Realbotix™'. In: *AI Love You: Developments in Human-Robot Intimate Relationships*. Ed. by Yuefang Zhou and Martin H. Fischer. Cham: Springer International Publishing, 2019, pp. 77–95. ISBN: 978-3-030-19734-6. DOI: 10.1007/978-3-030-19734-6.

[9] *Replika*. replika.com. URL: https://replika.com (visited on 09/03/2023).

[10] Kyle Wiggers and Amanda Silberling. *Meet Unstable Diffusion, the Group Trying to Monetize AI Porn Generators*. TechCrunch. 17th Nov. 2022. URL: https://techcrunch.com/2022/11/17/meet-unstable-diffusion-the-group-trying-to-monetize-ai-porn-generators/ (visited on 09/03/2023).

[11] Ben Dooley and Hisako Ueno. 'This Man Married a Fictional Character. He'd Like You to Hear Him Out.' In: *The New York Times*. *Business* (24th Apr. 2022). ISSN: 0362-4331. URL: https://www.nytimes.com/2022/04/24/business/akihiko-kondo-fictional-character-relationships.html (visited on 09/03/2023).

[12] Yuefang Zhou and Martin H. Fischer, eds. *AI Love You*. Potsdam, Germany: Springer, 2019. URL: https://link.springer.com/book/10.1007/978-3-030-19734-6 (visited on 27/02/2023).

[13] Adrian David Cheok, David Levy, Kasun Karunanayaka and Yukihiro Morisawa. 'Love and Sex with Robots'. In: *Handbook of Digital Games and Entertainment Technologies*. Ed. by Ryohei Nakatsu, Matthias Rauterberg and Paolo Ciancarini. Singapore: Springer Singapore, 2017, pp. 833–858. ISBN: 978-981-4560-49-8 978-981-4560-50-4. DOI: 10.1007/978-981-4560-50-4_15.

[14] Bedbible Research Center. *Sex Robot Industry: State of Market Size, Technology (AI), User Sentiment, and Other Statistics.* Bedbible.com. 24th Oct. 2022. URL: https://bedbible.com/sex-robot-industry-market-size-technology-ai-user-sentiment-statistics/ (visited on 09/03/2023).

[15] Bernard Marr. *Future Of Intimacy: Sex Bots, Virtual Reality, And Smart Sex Toys*. Forbes. URL: https://www.forbes.com/sites/bernardmarr/2020/11/30/future-of-intimacy-sex-bots-virtual-reality-and-smart-sex-toys/?sh=21727cfc38fa (visited on 09/03/2023).

[16] Masahiro Mori, Karl F. MacDorman and Norri Kageki. 'The Uncanny Valley [From the Field]'. In: *IEEE Robotics & Automation Magazine* 19.2 (June 2012), pp. 98–100. ISSN: 1558-223X. DOI: 10.1109/MRA.2012.2192811. https://ieeexplore.ieee.org/document/6213238

[17] Alexandra Sternlicht. *CarynAI Will Be Your Girlfriend for $1 a Minute*. Fortune. 5th Sept. 2023.
URL: https://fortune.com/2023/05/09/snapchat-influencer-launches-carynai-virtual-girlfriend-bot-openai-gpt4/ (visited on 27/07/2023).

[18] Matthias Scheutz and Thomas Arnold. 'Are We Ready for Sex Robots?' In: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, Mar. 2016, pp. 351–358. ISBN: 978-1-4673- 8370-7. DOI: 10.1109/HRI.2016.7451772.

[19] Jenna Owsianik. 'Sexbot Market Guide: The Best AI Sex Dolls and Sex Robots'. In: *Future of Sex* (10th Feb. 2022).
URL: https://futureofsex.net/robots/state-of-the-sexbot-market-the-worlds-best-sex-robot-and-ai-love-doll-companies/ (visited on 09/03/2023).

[20] Andrew R. Chow. *Why People Are Confessing Their Love For AI Chatbots.* Time. 23rd Feb. 2023. URL: https://time.com/6257790/ai-chatbots-love/ (visited on 09/03/2023).

[21] Ian Pearson. 'The Future of Sex Report: The Rise of the Robosexuals'. In: *Bondara* (2015).

[22] David Levy. *Love and Sex with Robots*. Gerald Duckworth & Co Ltd, 9th Apr. 2009. 352 pp. ISBN: 978-0-7156-3777-7.

[23] Amelia Fiske, Peter Henningsen and Alena Buyx. 'Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy'. In: *Journal of Medical Internet Research* 21.5 (9th May 2019), e13216. DOI: 10.2196/13216. URL: https://www.jmir.org/2019/5/e13216 (visited on 13/03/2023).

[24] Eduard Fosch-Villaronga and Adam Poulsen. 'Sex Care Robots: Exploring the Potential Use of Sexual Robot Technologies for Disabled and Elder Care'. In: *Paladyn, Journal of Behavioral Robotics* 11.1 (1st Jan. 2020), pp. 1–18. ISSN: 2081-4836. DOI: 10.1515/pjbr-2020-0001. URL: https://www.degruyter.com/document/doi/10.1515/pjbr-2020-0001/html?lang=en (visited on 10/03/2023).

[25] Charlotta Odlind and Kathleen Richardson. 'The End of Sex Robots—For the Dignity of Women and Girls'. In: *Man-Made Women: The Sexual Politics of Sex Dolls and Sex Robots*. Ed. by Kathleen Richardson and Charlotta Odlind. Social and Cultural Studies of Robots and AI. Cham: Springer International Publishing, 2022, pp. 1–16. ISBN: 978-3-031-19381-1. DOI: 10.1007/978-3-031- 19381-1_1. URL: https://doi.org/10.1007/978-3-031-19381-1_1 (visited on 09/03/2023).

[26] Shirley MacWilliam. 'Playthings and Corpses—Turning Women into Dead Body Objects: Sexual Objectification, Victimisation, Representation and Consent in Art and Sex Dolls/Robots'. In: *Man-Made Women: The Sexual Politics of Sex Dolls and Sex Robots*. Ed. by Kathleen Richardson and Charlotta Odlind. Social and Cultural Studies of Robots and AI. Cham: Springer International Publishing, 2022, pp. 67–89. ISBN: 978-3-031-19381-1. DOI: 10.1007/978-3-031-19381-1_5.

[27] *Outlaw Poverty Not Prostitutes*. English Collective of Prostitutes. 19th Mar. 2019. URL: https://prostitutescollective.net/outlaw-poverty-not-prostitutes/ (visited on 17/03/2023).

[28] Sherry Turkle. *Alone Together: Why We Expect More from Technology and Less from Each Other*. New York, NY, US: Basic Books, 2011, pp. xvii, 360. ISBN: 978-0-465-01021-9.

[29] Lydia Delicado-Moratalla. 'Mapping the Uses of 'Sex' Dolls: Pornographic Content, Doll Brothels and the Similarities with Rape'. In: *Man-Made Women: The Sexual Politics of Sex Dolls and Sex Robots*. Ed. by Kathleen Richardson and Charlotta Odlind. Social and Cultural Studies of Robots and AI. Cham: Springer International Publishing, 2022, pp. 35–47. ISBN: 978-3-031-19381-1. DOI: 10.1007/978-3-031-19381-1_3.

[30] Mark Migotti and Nicole Wyatt. 'On the Very Idea of Sex with Robots'. In: *Robot Sex: Social and Ethical Implications*. Cambridge, MA: MIT Press, 2017, pp. 31–46.

[31] John P. Sullins. 'Robots, Love, and Sex: The Ethics of Building a Love Machine'. In: IEEE Transactions on Affective Computing 3.4 (2012), pp. 398–409. ISSN: 1949-3045. DOI: 10.1109/T AFFC.2012.31.

[32] Oliver Bendel. 'Sex Robots from the Perspective of Machine Ethics'. In: *Love and Sex with Robots*. Ed. by Adrian David Cheok, Kate Devlin and David Levy. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 17–26. ISBN: 978-3-319-57738-8. DOI: 10.1007/978-3-319-57738-8_2.

[33] Adrienne de Ruiter. 'The Distinct Wrong of Deepfakes'. In: *Philosophy & Technology* 34.4 (1st Dec. 2021), pp. 1311–1332. ISSN: 2210-5441. DOI: 10 . 1007 / s13347 - 021 - 00459 - 2. URL: https://link.springer.com/article/10.1007/s13347-021-00459-2 (visited on 30/03/2023).

[34] Samantha Cole. *This AI Tool Is Being Used to Make Freaky, Machine-Generated Porn*. Vice. 24th Aug. 2022. URL: https://www.vice.com/en/article/xgygy4/stable-diffusion-stability-ai-nsfw-ai-generated-porn (visited on 30/03/2023).

[35] Samantha Cole, Emanuel Maiberg and Anna Koslerova. *'Frankenstein's Monster:' Images of Sexual Abuse Are Fueling Algorithmic Porn*. Vice. 10th Nov. 2020. URL: https://www.vice.com/en/article/akdgnp/sexual-abuse-fueling-ai-porn-deepfake-czech-casting-girls-do-porn (visited on 30/03/2023).

[36] Shannon Vallor and William J. Rewak. *An Introduction to Data Ethics*. 2018. URL: https://www.scu.edu/media/ethics-center/technology-ethics/IntroToDataEthics.pdf (visited on 04/03/2023).

[37] Haibo Zhang, Toru Nakamura, Takamasa Isohara and Kouichi Sakurai. 'A Review on Machine Unlearning'. In: *SN Computer Science* 4.4 (19th Apr. 2023), p. 337. ISSN: 2661-8907. DOI: 10. 1007/s42979-023-01767-4. URL: https://doi.org/10.1007/s42979-023-01767-4 (visited on 27/07/2023).

[38] Melissa Heikkilä. *The Viral AI Avatar App Lensa Undressed Me—without My Consent*. MIT Technology Review. 12th Dec. 2022. URL: https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/ (visited on 30/03/2023).

[39] Brian R. Duffy. 'Anthropomorphism and the Social Robot'. In: *Robotics and Autonomous Systems*. Socially Interactive Robots 42.3 (31st Mar. 2003), pp. 177–190. ISSN: 0921-8890. DOI: 10.1016/ S0921-8890(02)00374-3.

[40] Daniel T. Gilbert and Edward E. Jones. 'Perceiver-Induced Constraint: Interpretations of Self Generated Reality'. In: *Journal of Personality and Social Psychology* 50 (1986), pp. 269–280. ISSN: 1939-1315. DOI: 10.1037/0022-3514.50.2.269.

[41] Alistair M. C. Isaac and Will Bridewell. 'White Lies on Silver Tongues: Why Robots Need to Deceive (and How)'. In: *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. Ed. by Patrick Lin, Keith Abney and Ryan Jenkins. Oxford University Press, 30th Nov. 2017, p. 0. ISBN: 978-0-19- 065295-1. DOI: 10.1093/oso/9780190652951.003.0011. URL: https://academic.oup.com/book/2320/chapter-abstract/142468698?redirectedFrom=fulltext (visited on 13/03/2023).

[42] Sven Nyholm and Lily Eva Frank. 'It Loves Me, It Loves Me Not: Is It Morally Problematic to Design Sex Robots That Appear to Love Their Owners?' In: *Techné: Research in Philosophy and Technology* (6th Dec. 2019). DOI: 10.5840/techne2019122110. URL: https://www.pdcnet.org/pdc/bvdb.nsf/purchase?openform&fp=techne&id=techne_2019_0999_12_2_110 (visited on 13/03/2023).
[43] Kate Devlin. 'The Ethics of the Artificial Lover'. In: *The Ethics of Artificial Intelligence*. Oxford University Press, 2019, pp. 271–290.

[44] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen and Abraham Bernstein. 'Implementations in Machine Ethics: A Survey'. In: *ACM Computing Surveys* 53.6 (30th Nov. 2021), pp. 1–38. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3419633. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4364029 (visited on 13/03/2023).

[45] Kevin Roose. 'A Conversation With Bing's Chatbot Left Me Deeply Unsettled'. In: *The New York Times. Technology* (16th Feb. 2023). ISSN: 0362-4331. URL: https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html (visited on 13/03/2023).

[46] James Vincent. 'Twitter Taught Microsoft's AI Chatbot to Be a Racist Asshole in Less than a Day'. In: *The Verge* (24th Mar. 2016). URL: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist (visited on 14/03/2023).

[47] Yuntao Bai et al. *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*. 12th Apr. 2022. DOI: 10.48550/arXiv.2204.05862. arXiv: 2204.05862 [cs]. URL: http://arxiv.org/abs/2204.05862 (visited on 28/03/2023). preprint.

[48] Jinying Lin, Zhen Ma, Randy Gomez, Keisuke Nakamura, Bo He and Guangliang Li. 'A Review on Interactive Reinforcement Learning From Human Social Feedback'. In: *IEEE Access 8* (2020), pp. 120757–120765. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.3006254.

[49] Jeannie Suk Gersen. 'Sex Lex Machina: Intimacy and Artificial Intelligence'. In: *Columbia Law Review* 119.7 (2019), pp. 1793–1810. ISSN: 0010-1958. JSTOR: 26810849. URL: https://www.jstor.org/stable/26810849 (visited on 10/03/2023).

[50] Francis X. Shen. 'Sex Robots Are Here, but Laws Aren't Keeping up with the Ethical and Privacy Issues They Raise'. In: *The Conversation* (12th Feb. 2019). URL: https://theconversation.com/sex-robots-are-here-but-laws-arent-keeping-up-with-the-ethical-and-privacyissues-they-raise-109852 (visited on 14/03/2023).

[51] Beatriz Yumi Aoki and Takeshi Kimura. 'Sexuality and Affection in the Time of Technological Innovation: Artificial Partners in the Japanese Context'. In: *Religions* 12.5 (5 May 2021), p. 296. ISSN: 2077-1444. DOI: 10.3390/rel12050296. URL: https://www.mdpi.com/2077-1444/12/5/296 (visited on 10/03/2023).

[52] *TROTTLA*. 2011. URL: http://www.trottla.net/eutop.html (visited on 27/03/2023).

[53] Roc Morin. 'Can Child Dolls Keep Pedophiles from Offending?' In: *The Atlantic. Health* (11th Jan. 2016). URL: https://www.theatlantic.com/health/archive/2016/01/can-child-dolls-keep-pedophiles-from-offending/423324/ (visited on 14/03/2023).

[54] Yusuff Jelili Amuda and Ismaila B. Tijani. *Ethical and Legal Implications of Sex Robot: An Islamic Perspective*. 19th Feb. 2012. URL: https://papers.ssrn.com/abstract=2008011 (visited on 09/03/2023). preprint.

[55] Sonya Ziaja. 'Homewrecker 2.0: An Exploration of Liability for Heart Balm Torts Involving AI Humanoid Consorts'. In: *Social Robotics*. Ed. by Bilge Mutlu, Christoph Bartneck, Jaap Ham, Vanessa Evers and Takayuki Kanda. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2011, pp. 114–124. ISBN: 978-3-642-25504-5. DOI: 10.1007/978-3-642-25504-5_12.

[56] Anna C. B. Russell. 'Blurring the Love Lines: The Legal Implications of Intimacy with Machines'. In: *Computer Law & Security Review* 25.5 (1st Sept. 2009), pp. 455–463. ISSN: 0267-3649. DOI: 10.1016/j.clsr.2009.07.003. URL: https://www.sciencedirect.com/science/article/pii/ S0267364909001253 (visited on 09/03/2023).

[57] Matthew Rueben, Alexander Mois Aroyo, Christoph Lutz, Johannes Schmölz, Pieter Van Cleynenbreugel, Andrea Corti, Siddharth Agrawal and William D. Smart. 'Themes and Research Directions in Privacy-Sensitive Robotics'. In: *2018 IEEE Workshop on Advanced Robotics and Its Social Impacts* (ARSO). Sept. 2018, pp. 77–84. DOI: 10.1109/ARSO.2018.8625758.

[58] John Danaher. 'Robotic Rape and Robotic Child Sexual Abuse: Should They Be Criminalised?' In: *Criminal Law and Philosophy* 11.1 (1st Mar. 2017), pp. 71–95. ISSN: 1871-9805. DOI: 10. 1007/s11572-014-9362-x. URL: https://doi.org/10.1007/s11572-014-9362-x (visited on 14/03/2023).

How AI Researchers Are Redirecting AI's Societal Impact

Priyanka Nanayakkara, PhD Candidate
Computer Science & Communication, Northwestern University
Evanston, Illinois, United States

ABSTRACT
Over the past few years, the AI research community has embarked on an internal reckoning around the impacts of its work in ways that are reshaping the research process and could have profound impacts on the direction of AI more broadly. Reflecting on efforts to increase ethical deliberation within the field (e.g., the NeurIPS "broader impact statement") is paramount to understanding how AI researchers are making decisions today that will influence the impact of AI over the next several decades. The 2026 Report comes at an ideal time for such reflection: Ethics-related efforts are underway, but still shifting in important ways. Hence, the 2026 Report presents an opportunity for documenting a time of crucial change.

The 2021 AI100 report [1] includes significantly more discussion of negative impacts than the 2016 Report [2], illustrating an "inflection point" [1] in AI conversations. While the 2016 Report takes an optimistic tone, describing how AI can impact life in a North American city from transportation to entertainment, the 2021 Report is more somber, with significant discussion of negative consequences. As the public's awareness around AI (and its pitfalls) has increased, the AI research community has begun an internal reckoning around the impacts of its work in ways that are reshaping the research process and over time could have profound impacts on the direction of AI development and application. Hence, these shifts will be crucial to engage with in the 2026 Report.

In particular, the AI research community has started to implement community-wide mechanisms for encouraging ethical deliberation over potential implications of research contributions. For example, in 2020 a top machine learning conference, NeurIPS, required all authors to write a "broader impact" statement on potential positive *and* negative consequences of their work [3]; statements were published alongside research. This practice is relatively nascent for the field, but is likely to gain further importance given (1) AI's increasing capacities and pervasiveness in everyday life and (2) AI researchers' unique position to make upstream changes that could influence downstream applications and society at large. The longitudinal nature of the AI100 reports is an ideal form for studying field-wide efforts, as such changes may be incremental, but potentially far-reaching.

Already, there is a growing body of research around the broader impact statement and other ethics-related efforts at the research level (e.g., [4]–[14]). In 2020, I led a study [10] analyzing the first NeurIPS broader impact statements. Through a thematic analysis of 300

statements, my collaborators and I identified several themes describing topics authors discussed and how they discussed these topics (for example, in terms of specificity and valence of anticipated consequences). There is significant overlap between topics discussed across the 2016 and 2021 reports and 2020 statements. We found that authors focused on societally oriented impacts to areas like privacy, bias, labor, environment, and media. Authors also described more technically oriented impacts, such as implications around robustness and reliability, accuracy, and generalizability, suggesting that giving authors guidance for mapping more technically oriented consequences to societal impacts may help broaden the space of identified consequences. In almost 10% of our sample, authors implied that their theoretical work did *not* have foreseeable societal consequences, raising questions about the extent to which the impact statement is an effective mechanism for steering the direction of non-applied work. Similarly, Ashurst et al. [8] found that about 10% of papers chose to "opt out" of the statement by saying it was "not applicable" to the work at hand. We found that other times, authors implied that the computer science research community is responsible for mitigating negative consequences, meaning that broader impact statements could serve as useful guides for identifying future research areas.

In 2021, NeurIPS took a slightly different approach to encouraging ethical deliberation: (1) an impact statement became optional while a "paper checklist" was required [15], (2) authors were explicitly encouraged to discuss *negative* consequences and mitigation strategies as per the checklist's guidelines, and (3) papers could now be rejected on ethical grounds. I co-led a study with David Liu [11] investigating outcomes of the updated ethics requirements. Our collaborators and we qualitatively analyzed over 200 impact statements and all available ethics reviews. We found that authors seemed to express a lack of agency around identifying and mitigating negative societal consequences, citing adversarial users and misuses that are difficult to predict at the research stage, and that authors rarely proposed mitigation strategies.

We found that authors focused on societally-oriented impacts to areas like privacy, bias, labor, environment, and media. Authors also described more technically-oriented impacts, such as implications around robustness & reliability, accuracy, and generalizability, suggesting that giving authors guidance for mapping more technically-oriented consequences to societal impacts may help broaden the space of identified consequences.

However, our analysis of ethics reviews indicated that authors were often willing to accept suggestions from ethics reviewers. 2021's explicit instructions to focus on negative consequences may have helped expose ways in which AI researchers on their own may struggle with foreseeing consequences (which is often difficult if not impossible) but in collaboration with other experts, such as ethics reviewers, may be able to make further inroads in redirecting AI research. Such collaboration may also help overcome the "[t]rivialisation of ethics and governance," which Prunkl et al. [7] suggest can occur if impact statements lead researchers to "form the impression that it is possible to fully anticipate the ethical

and societal consequences of one's research in such a statement, thereby trivializing the complexity of the task and the efforts needed." Regardless, future iterations of NeurIPS and other conferences implementing ethics-related processes will help illuminate effective strategies.

The 2021 Report [1], published shortly after NeurIPS' first impact statement requirement, makes brief mention that such statements are being implemented across "several prominent AI conferences," but goes no further. The 2026 Report can therefore pick up where the 2021 Report [2] left off by attempting answers to questions about the effectiveness of ethics-related efforts: *Are consequences researchers identify changing over time? To what extent is there an increase in research around mitigating negative consequences identified in previous years' impact statements and checklists? How have researchers responded to each iteration of the impact statement over the past seven years, and what are some challenges they've faced? How are impact statements generally shaping AI research?*

By 2026, we will have at least three more years of evidence around the AI research community's efforts, including future iterations of impact statements as well as other efforts, such as the newly formed Ethics and Society Review Board [13], which reviews research proposals prior to funding. Furthermore, the study of these efforts is an active research area that will undoubtedly continue to yield learnings from each new attempt to facilitate ethical deliberation. Investigating how AI researchers are making decisions today related to their works' future societal consequences is paramount to understanding AI and its impacts on society over the next several decades. The 2026 Report therefore comes at an ideal time where ethics-related efforts are already underway, but still shifting in important ways.

Thank you to Nicholas Diakopoulos for comments on a draft.

References

[1] M. L. Littman, I. Ajunwa, G. Berger, C. Boutilier, M. Currie, F. Doshi-Velez, G. Hadfield, M. C. Horowitz, C. Isbell, H. Kitano, K. Levy, T. Lyons, M. Mitchell, S. Julie, S. Sloman, S. Vallor, and T. Walsh, "Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report," *Stanford University*, 2021. [Online]. Available: http://ai100.stanford.edu/2021-report.

[2] P. Stone, R. Brooks, E. Brynjolfsson, R. Calo, O. Etzioni, G. Hager, J. Hirschberg, S. Kalyanakrishnan, E. Kamar, S. Kraus, K. Leyton-Brown, D. Parkes, W. Press, A. Saxenian, J. Shah, M. Tambe, and A. Teller, "'Artificial Intelligence and Life in 2030.' One Hundred Year Study on Artificial Intelligence," *Stanford University*, 2016. [Online]. Available: https://ai100.stanford.edu/2016-report.

[3] H.-T. Lin, M.-F. Balcan, R. Hadsell, and M. Ranzato, *NeurIPS 2020 Call for Papers, 2020*. [Online]. Available: https://neurips.cc/Conferences/2020/CallForPapers.

[4] B. Hecht, L. Wilcox, J. P. Bigham, J. Schöning, E. Hoque, J. Ernst, Y. Bisk, L. De Russis, L. Yarosh, B. Anjum, D. Contractor, and C. Wu, "It's time to do something: Mitigating the negative impacts of computing through a change to the peer review process," *arXiv preprint arXiv:2112.09544*, 2021.

[5] K. Sim, A. Brown, and A. Hassoun, "Thinking Through and Writing About Research Ethics Beyond 'Broader Impact'," *ACM FAccT Tutorial*, 2021.

[6] P. Nanayakkara, N. Diakopoulos, and J. Hullman, "Anticipatory Ethics and the Role of Uncertainty," *Navigating the Broader Impacts of AI Research Workshop, NeurIPS*, 2020.

[7] C. E. Prunkl, C. Ashurst, M. Anderljung, H. Webb, J. Leike, and A. Dafoe, "Institutionalizing ethics in AI through broader impact requirements," *Nature Machine Intelligence*, vol. 3, no. 2, pp. 104–110, 2021.

[8] C. Ashurst, E. Hine, P. Sedille, and A. Carlier, "AI Ethics Statements: Analysis and Lessons Learnt from NeurIPS Broader Impact Statements," in *ACM FAccT*, 2022, pp. 2047–2056.

[9] G. Abuhamad and C. Rheault, "Like a Researcher Stating Broader Impact For the Very First Time," *Navigating the Broader Impacts of AI Research Workshop*, NeurIPS, 2020.

[10] P. Nanayakkara, J. Hullman, and N. Diakopoulos, "Unpacking the Expressed Consequences of AI Research in Broader Impact Statements," in *AAAI/ACM AIES*, 2021, pp. 795–806.

[11] D. Liu, P. Nanayakkara, S. A. Sakha, G. Abuhamad, S. L. Blodgett, N. Diakopoulos, J. Hullman, and T. Eliassi-Rad, "Examining Responsibility and Deliberation in AI Impact Statements and Ethics Reviews," in *AAAI/ACM AIES*, 2022, pp. 424–435.

[12] M. Boyarskaya, A. Olteanu, and K. Crawford, "Overcoming Failures of Imagination in AI Infused System Development and Deployment," *Navigating the Broader Impacts of AI Research Workshop, NeurIPS*, 2020.

[13] M.S. Bernstein, M. Levi, D. Magnus, B.A. Rajala, D. Satz, and Q. Waeiss, "Ethics and society review of Artificial Intelligence Research," *arXiv preprint arXiv:2106.11521*, 2021.

[14] C. Ashurst, S. Barocas, R. Campbell, and D. Raji, "Disentangling the Components of Ethical Research in Machine Learning," in *ACM FAccT*, 2022, pp. 2057–2068.

[15] NeurIPS, *NeurIPS 2021 Paper Checklist Guidelines*, https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist, 2021.

Reconsidering Interaction Between AI and Religion According to the AI100 Reports

Dr. Ximian Xu, Postdoctoral Research Fellow
Theology & Ethics of Artificial Intelligence, University of Edinburgh
Edinburgh, Scotland

ABSTRACT
The AI100 Study Panel reports offer guidelines both for AI research and interdisciplinary studies on AI. However, AI-and-religion studies, which have been thriving for three decades, are omitted throughout the two reports. Does the Study Panel make no contributions to religious engagement with AI? The fact is that the two reports do provide two guiding principles for AI-and-religion research in the near term. First, AI's complementary strengths remind us to investigate how religious and theological ethics can address ethical issues surrounding AI's augmentation of human activity and ability. Second, placing emphasis on normativity as a grand challenge to AI, the reports force us to consider the normativity of different religious systems while deploying AI-driven devices within specific religious communities. These two principles should be further developed by the time of the next Study Panel report so that AI-and-religion studies can enhance the interdisciplinary feature of AI100.

Artificial intelligence (AI) is pervasive in human lives in the twenty-first century. Against this backdrop, AI100 was launched to explore the field of AI and look into the intertwinement between AI and humans.[1] The two AI100 Study Panel reports (2016 and 2021) have brought to light both the rapid progress of AI technology and AI's extensive influences at numerous levels of human lives as well as communities. However, the interaction between AI and religion falls through the cracks in both reports. Nonetheless, I will demonstrate that the reports provide two guiding principles for steering the course of AI-and-religion studies in the near term.

The past three decades saw the rise and flourishing of AI-and-religion studies.[2] The development of AI-and-religion studies is characterized by the fact that much scholarly attention was and continues to be drawn to religious engagement with fictional AI, that is, artificial general intelligence (AGI) and artificial superintelligence (ASI). Some scholars dive into religious ideas behind

1 Peter Stone, Rodney Brooks, Erik Brynjolfsson et al., *Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence 2016 Study Panel Report*, Stanford University (Stanford CA, September 2016), 1; hereafter *Report 2016*, https://ai100.stanford.edu/2016-report; Michael L. Littman, Ifeoma Ajunwa, Guy Berger et al., *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report*, Stanford University (Stanford, CA, September 2021), 1–2; hereafter *Report 2021*, http://ai100.stanford.edu/2021-report.

2 Two classic and groudbreaking works are Noreen L. Herzfeld, *In Our Image: Artificial Intelligence and the Human Spirit* (Minneapolis, MN: Fortress, 2002); Anne Foerst, *God in the Machine: What Robots Teach Us about Humanity and God* (New York: Plume, 2005).

AGI [3] and ASI.[4] Others painstakingly demonstrate that religion can contribute to the development of safe AGI and ASI. In some studies, AI and transhumanism are conjoined to illustrate technological Singularity, which refers to the coming age where intelligent machines overwhelm humans.[5] Indeed, some recent studies investigate the interaction between AI and religion with an eye to addressing ethical issues surrounding the application of AI in human life.[6] Be that as it may, some theologians and religious studies scholars seek to deal with ethical issues with particular attention to AGI, exploring how religion can help design virtuous and moral AGI.[7] As such, the present literature on AI-and-religion research is, by and large, dominated by fictional AI.

Both AI100 reports bracket off religious and theological inquiry into AI and do not pay due attention to how AI may have a bearing on religious communities. Despite this omission, the reports do offer two guiding principles for AI-and-religion research in the near term.

First, the two reports pave a way for future AI-and-religion research by flagging up AI's complementary strengths. In the first report, the Study Panel observes that research interests are growing in how humans enable "AI systems to overcome their limitations" and how AI can "augment human abilities and activities."[8] In the second report, the Study Panel fleshes out and underscores AI's complementary strengths to '[augment] human capabilities' as one of the most promising opportunities for AI. This AI-based augmentation includes, among others, AI-assisted insights, AI-assisted decision-making, and AI-based tools.[9]

The emphasis on AI's complementary strengths can help us adjust AI-and-religion research to reveal the contributions that religion and theology can make to AI ethics.

The emphasis on AI's complementary strengths can help us adjust AI-and-religion research to reveal the contributions that religion and theology can make to AI ethics. As Mark Coeckelbergh notes, "AI ethics is about technological change and its impact on individual lives."[10] Rather than being obsessed with fictional AI, religious studies scholars and theologians should uncover how theological and religious resources can assist in addressing ethical issues surrounding AI's complementary strengths actualized in human life. A few scholars have taken a step in this direction. For example, Amy Michelle DeBaets

3 See, for example, Robert M. Geraci, *Apocalyptic AI: Visions of Heaven in Robotics, Artificial Intelligence, and Virtual Reality* (Oxford: Oxford University Press, 2010).

4 See, for example, Yong Sup Song, "Religious AI as an Option to the Risks of Superintelligence: A Protestant Theological Perspective," *Theology and Science* 19, no. 1 (2021): 65–78; Ilia Delio, *Re-Enchanting the Earth: Why AI Needs Religion* (Maryknoll, NY: Orbis Books, 2020).

5 See, for example, Calvin Mercer and Tracy J. Trothen, *Religion and the Technological Future: An Introduction to Biohacking, Artificial Intelligence, and Transhumanism* (Cham: Palgrave Macmillan, 2021).

6 See, for example, John Wyatt and Stephen N. Williams, eds., *The Robot Will See You Now: Artificial Intelligence and the Christian Faith* (London: SPCK, 2021); Benedikt Paul Göcke and Astrid Rosenthal-von der Pütten, eds., *Artificial Intelligence: Reflections in Philosophy, Theology, and the Social Sciences* (Paderborn: Mentis, 2020).

7 A latest example of this type is Jaco J. Hamman, *Pastoral Virtues for Artificial Intelligence: Care and the Algorithms that Guide Our Lives* (Lanham: Lexington Books, 2022).

8 *Report* 2016, 17.

9 *Report* 2021, 48–51.

10 Mark Coeckelbergh, *AI Ethics* (Cambridge, Massachusetts: The MIT Press, 2020), 9.

elucidates the way in which the Christian idea of love can help deal with the ethical questions about carebots and AI-enhanced healthcare.[11] Yet, there is still a lot to work out on this front. Hence, the two reports open up vistas for the formation and thriving of theological/religious AI ethics.

Second, touching upon normativity as a grand challenge to AI, the second report reminds religious communities that normativity is an essential principle of AI-and-religion research. The Study Panel rightly asserts that unanimous normativity for AI cannot be achieved insofar as human normative systems vary across human cultures.[12] For this reason, ethical AI 'needs to have good normative models and to be capable of integrating its behavior into human normative institutions and processes.'[13]

To be sure, religions vary from each other, and even communities of one religion have different traditions and communal norms. It is hardly convinced that an AI system can be designed for all religions with universally unanimous religious values and norms. In this light, the second report stimulates interdisciplinary studies on AI and religion to move from *religion*-and-AI to *religions*-and-AI so as to attend to various religious norms for AI. But how?

The second report shows a path to making such a scholarly breakthrough. It spends some space discussing that AI-enhanced care will have lasting impacts on human-to-human care and 'reshape traditional caring relationships.'[14] New caring relationships give us the inspiration that religious pastoral care can serve as a hinge to join specific religious norms and AI. Religious normativity is tied up with religious life, of which religious pastoral care is an indispensable component. Hence, AI-enhanced religious pastoral care can lead to an AI that is capable of integrating its behavior into a normative religious community so as to foster the growth of that community.

Advancing *religions*-and-AI through exploration of AI-enhanced religious pastoral care will conduce to the closer cooperation between religious communities and the AI industry. That is, *religions*-and-AI studies can introduce various religious norms and values to AI research so that AI designers can know the particularities of religions. An example of this type is the Church of England Alexa Skill. Since 2018, Alexa has been used among the congregations of the Church of England. The Alexa Skill provides, among others, "say a prayer," "explore the Christian faith," and "mental health reflection."[15] Yet, AI-enhanced Christian pastoral care, which is linked to and varies between local communities, remains to be developed and programmed into the Alexa Skill. It can be anticipated that *religions*-and-AI studies will foster the collaboration between religions and AI companies in designing AI systems for particular religious communities.

The two AI100 Study Panel reports have afforded much guidance for interdisciplinary studies on AI. Notwithstanding the Study Panel's omission of religion, it holds true that religious and theological engagement with AI can benefit much from the two reports. Above all, AI's complementary strengths and AI for a normative religious community will facilitate the development of *religions*-and-AI research such as to transform AI-and-religion studies by the time of the next report.

11 Amy Michelle DeBaets, "The Robot Will See You Now: Reflections on Technologies in Healthcare," in *Love, Technology and Theology*, ed. Scott A. Midson (London: T&T Clark, 2020), 93–108.

12 *Report 2021*, 23.

13 Ibid., 24.

14 Ibid., 68.

15 'Church of England Alexa skill asked 75,000 questions in first year,' accessed 25 January 2023, https://livingchurch.org/2019/05/30/alexa-ask-the-church-of-england/#:~:text=Embrace%20of%20social%20media%20by,name%20for%20voice%2Dactivated%20service