UNIVERSITY OF COPENHAGEN
FACULTY OF SCIENCE

Blackness is a culture and an experience.
How do you feel about being queer?
I'm queer and I love it!
Because they cant bear the pain.
Addicts can't bear the pain of what?
Of existing.

THE DATA IS OF US BUT NOT US

Exploring Relations to Responsive AI Systems Through Speculative Participatory Art-Based Research

Mirabelle Jones

Supervisors:
Irina Shklovski (University of Copenhagen)
Christina Neumayer (University of Copenhagen)
Jakob Grue Simonsen (University of Copenhagen)

Submitted: July 31, 2024
This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen.

Abstract

As artificial intelligence becomes increasingly embedded into society, invisibly and visibly intertwined in people's daily lives and identities, it becomes crucial to consider the ethical implications of the resulting technological systems. While several researchers in HCI have pursued mitigating the potential harms and risks related to AI systems from a technological perspective through the development of AI and ML tools, others have considered these issues from a socio-technical lens, addressing the complexities of human-AI relations as deeply situated in the social, calling for interdisciplinary approaches. However, much of this research fails to consider knowledge formation as situated and embodied, discarding the importance of emotional, cognitive, and aesthetic knowledge as valuable findings in human-AI research. With the latter objective in mind, this dissertation explores human-AI relations as construed through how humans make sense of responsive AI systems by implementing an art-based participatory approach that highlights the importance of situated knowledge. I propose SPAR (Speculative Participatory Art-Based Research) as a method to further understandings regarding how people relate to emerging AI technologies through lived and felt experiences. SPAR adopts an inductive process of speculation, making, and scaffolding participatory artworks. I demonstrate this method in three speculative participatory art-based research projects focusing on generative AI systems that respond to and mimic human identities. In three papers and an annotated portfolio, I demonstrate the impact of: large language models on bodies using performance art, encounters with deep fakes of speculative alternative selves, and normativity and personalization in fine-tuned large language models. I conclude by speculating on the method's application towards future research in an emerging fourth wave of HCI.

Resumé

Efterhånden som kunstig intelligens i stigende grad bliver integreret i vores samfund, usynligt og synligt sammenflettet i menneskers daglige liv og identiteter, bliver det afgørende at overveje de etiske implikationer af de resulterende teknologiske systemer. Mens adskillige forskere indenfor HCI har forsøgt at afbløde de potentielle skader og risici forbundet med AI-systemer fra et teknologisk perspektiv gennem udvikling af AI- og ML-værktøjer, har andre overvejet disse spørgsmål fra en socio-teknisk vinkel og adresseret kompleksiteten af menneske-AI-relationer som dybt placeret i kaldet for tværfaglige tilgange. Men meget af denne forskning undlader at betragte vidensdannelse som situeret og legemliggjort, og glemmer derfor vigtigheden af emotionel, kognitiv og æstetisk viden som værdifulde resultater i human-AI-forskning. Med sidstnævnte mål i tankerne, udforsker denne afhandling menneske-AI-relationer som fortolket gennem den måde mennesker forstår AI-systemer ved at implementere en kunst- og deltagerbaseret tilgang, der fremhæver vigtigheden af situeret viden. Jeg foreslår SPAR (Speculative Participatory Art-Based Research) som en metode til at fremme forståelsen af, hvordan mennesker forholder sig til nye AI-teknologier gennem levede og følte oplevelser. SPAR anvender en induktiv proces med spekulation, fremstilling og rammesætning af kunstværker, der opstår gennem deltagelse. Jeg demonstrerer denne metode i tre spekulative kunst- og deltagerbaserede forskningsprojekter med fokus på generative AI-systemer, der reagerer på og efterligner menneskelige identiteter. I tre artikler og en kommenteret portefølje demonstrerer jeg virkningen af: large language models på kroppe, der bruger performancekunst, møder med deep fakes af spekulative alternative jeger og normativitet og personalisering i finjusterede store sprogmodeller. Jeg afslutter med at spekulere i metodens anvendelse til fremtidig forskning i en kommende fjerde bølge af HCI.

Acknowledgments

This dissertation combines interdisciplinary approaches to knowledge formation in HCI to investigate through speculative participatory art how humans relate to emerging technologies in Artificial Intelligence. Herein, I demonstrate the evolution of an exploratory method across three quite unusual projects in both art and technology contexts. As a result, the projects presented, as well as their findings, would not have been possible without the support and assistance of many individuals from a variety of domains, including researchers, artists, curators, designers, makers, writers, and developers (just to name a few).

I begin by thanking my supervisors: Irina Shklovski, Christina Neumayer, and Jakob Grue Simonsen from the Faculty of Science at the University of Copenhagen. Without your support, especially in this dissertation's initial and final stages, I do not think I could have brought it to completion. I want to also thank my co-authors Irina Shklovski, Chistina Neumayer, and Nastasia Griffioen for helping me structure compelling concepts as research outputs. I would also like to extend my deep appreciation to my co-authors on publications not included in this dissertation Daniela Rosner, Brett Halperin, Miguel Sicart, Kirsto Reitan Andersen, and Obaida Hanteer. I am grateful to metaLAB Harvard and especially Sarah Newman, Kim Albrecht, Lins Derry, Magda Romanska, Jeffrey Schnapp, Jeanette Andrews, as well as the Berkman Klein Center for engaging me in conversations and inspirational research during and beyond my change of research environment. Thank you to Tonia Sideri for welcoming me at Novo Nordisk’s AI Center of Excellence to assist in establishing AI Ethics practices through artistic and qualitative research in the pharmaceutical industry.

My colleagues in the Department of Computer Science at the University of Copenhagen have supported me through the many speed bumps of life that interrupt research, be it through long and engaging conversations, emotional support, ongoing encouragement, or suggestions for literature. I would like to particularly thank Valeria Borsotti, Athina Panotopoulou, Sonja Rattay, Natalia Avlona, Kellie Dunn, Trine Rask Nielsen, Bhaskar Dutt, Kristin Kaltenhäser, Mantas Cibulskis, Sarah Homewood, Tor-Salve Dalsgaard, Hubert Zając, Valkyrie Savage, Arpit Bhatia, and Barry Alan Brown as well as former colleagues Carlos E. Tejada, Hasti Stefi, Jonas Schjerlund, Andreea Muresan, Hyunyoung Kim, Thomas van Gemert and Martin Maunsbach.

Making sense of the conceptual mess that sometimes results from interdisciplinary methods would not be possible without friends and colleagues in the arts who gave me opportunities to grow my ideas into realities that I could extend into the world. Many of the projects I discuss in this dissertation required the support of curator, speaker, researcher, and writer Majken Overgaard who granted me the opportunity to freely explore and develop layers of participation in art and technology through deconstructing barriers between workshops, exhibitions, and research practices during the COVID-19 pandemic. I have also been inspired and supported by Kristoffer Ørum who has made me feel at home in exploring odd concepts through even odder materials here in Copenhagen. I would also like to thank Kirsten Dencker Hansen who has been an endless source of support, a strong and inspiring feminist, activist, artist and maker. I am grateful to the artists who I can happily call friends that have been essential to my recent creative endeavors including Gorjeoux Moon, Leslie Foster, Heather Hoxsey, Chelsea Coon, Emmett Palaima, Marissa Lynn, Rine Rodin, Chigozie Nri, Marcela Lucatelli and Maru Mushtrieva.

My previous colleagues and friends in other parts of the world have also supported emotionally and mentally including Anouk Wipprecht, Jade Seven, Drew Trajillo, Zarineh Agnew, Kyle Chan, Crutcher Dunnavant, Amy Lux, Genevieve Danger Berrick. I have been supported in ways I cannot express my gratitude enough by Cathy Laughlin and Tristan Crane.

I would like to dedicate this dissertation to my mother Debra Stevens and my father Steven Danzer who have encouraged me throughout the many twists and turns in my life to pursue my curiosity and creativity in every endeavor I undertake.

Publications

Included in this dissertation

Paper I.

Jones M, Neumayer C and Shklovski I (2023) "Embodying the Algorithm: Exploring Relationships with Large Language Models Through Artistic Performance." *In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 19 April 2023, pp. 1–24. CHI '23. Association for Computing Machinery. https://doi.org/10.1145/3544548.3580885

Paper II.

Jones M, Neumayer C, and Shklovski I (2023) "'It's Time We Talked': interactions with the self through personalized deepfakes." Submitted to New Media and Society. July 18, 2024.

Paper III.

Jones M, Griffieon N, Neumayer C, Shklovski I (2024) "Artificial Intimacy: Exploring normativity and personalization through fine-tuning LLM chatbots." Targeted for submission ACM Journal of Responsible Computing.

Not included in this dissertation

B. Halperin, M. Jones, D. Rosner, "Haunted Aesthetics and Otherworldly Possibilities: Generating (Dis)embodied Performance Videos with AI" DIS '23 Workshop Position Paper July 2023

M. Jones, N. Griffioen, I. Shklovski, O. Hanteer, "Artificial Intimacy: An Exploration of the Personal and Intimate in Natural Language Processing Models" *NordiCHI '22: Adjunct Proceedings of the 2022 Nordic Human-Computer Interaction Conference October 2022* https://doi.org/10.1145/3547522.3547719

M. Sicart, I. Shklovski, and M. Jones, "Can Machine Learning be Moral?" *Presentation at the Human Centered AI workshop at NeurIPS, Dec. 13, 2021*. doi: 10.48550/arXiv.2201.06921

K. Andersen, M. Overgaard, M. Jones, and I. Shklovski, "Collectively Reimagining Technology." *Morals and Machines, no. 2*, pp. 86-93, 2021

Preface

Towards the Limitations of Computer-Human Relations with Responsive Systems

The foundation for the projects presented in this thesis is a body of interactive artwork exploring the relations people form with responsive electronics and sensory systems, particularly human-computer experiences that use language, performance, and storytelling. In 2013, I completed my MFA in Book Art and Creative Writing at Mills College in Oakland, California, USA, with the project *Book Reads You*, a Kinect-based sculpture of an illuminated book that would audibly read people's gestures to produce a story based on the movements of their bodies. While most participants enjoyed the project due to its novelty, many commented that they were disappointed because they wanted a more expansive and responsive vocabulary from the program and visualizations of their movements and the text produced. They found the story told too literal and lacking in the kind of noetic creativity humans are capable of. This was my first introduction to the possibilities and limitations of computer-based interactive objects and initiated my artistic interrogation into how interactive algorithmic-based experiences could create specific individualized relations with people.

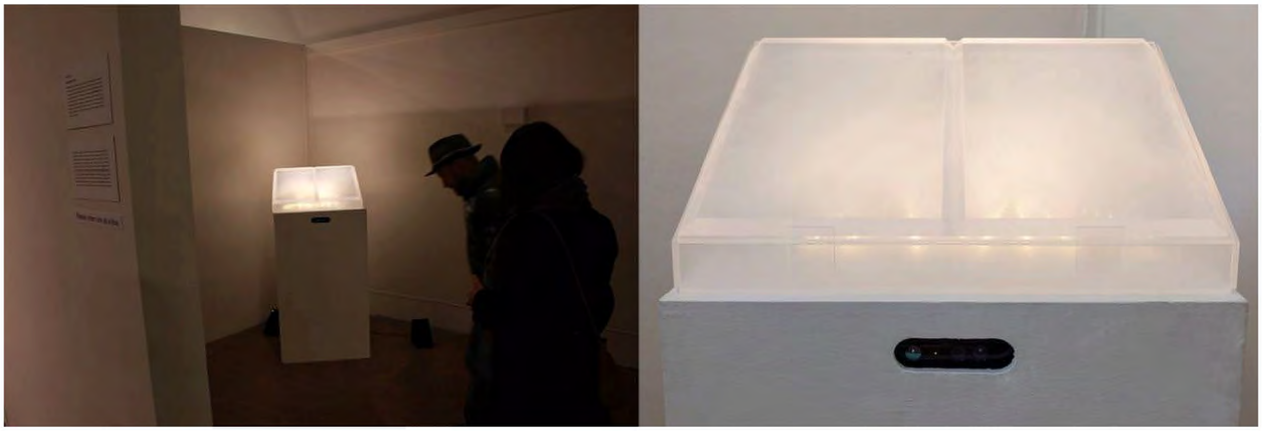
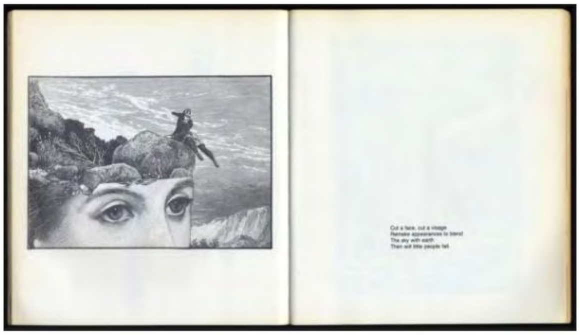
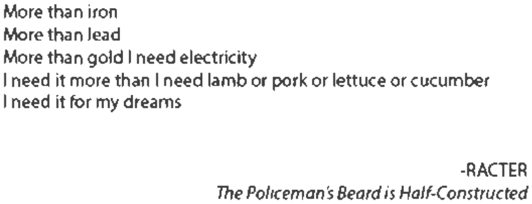


Figure 1 Book Reads You (2013) Mirabelle Jones at Swarm Gallery in Oakland, CA, USA. Image from MirabelleJones.com

In the written portion of my MFA thesis concerning the project, I explored the limitations of technologies to be used as adaptive actors, or as I put it at the time, their ability to be "creative." I approached the subject from a philosophy of mind perspective heavily influenced by Searle (1984) and a rejection of dualism, considering human-computer interaction as grounded in discussions of the body while simultaneously examining the roots of "creative AI." To the latter, I looked at early uses of chatbots such as RACTER, a program written by William Chamberlain who published what he reported as, "the first book of poetry 'written' by a computer" titled *The Policeman's Beard is Half-Constructed* (RACTER, 1984). I ultimately concluded that computers are only capable of mimicking creativity through randomization. I also recognized that this randomization sometimes leads to "slip-ups" such as repetition, nonsequiturs, and other linguistic oddities that can immediately identify to the participant that they are interacting with a machine incapable of telling their story.





More than iron
More than lead
More than gold I need electricity
I need it more than I need lamb or pork or lettuce or cucumber
I need it for my dreams

-RACTER
The Policeman's Beard is Half-Constructed

2. (Left): Pages from *The Policeman's Beard is Half-Constructed* with art by Joan Hall (RACTER, 1984)

3. (Right) Poem quoted from *The Policeman's Beard is Half-Constructed* (RACTER, 1984)

After serving as a Designer and Developer for the interactive arts organization Meow Wolf, where I aided in developing building-scale interactive experiences that would be visited by thousands of people a day, I moved to Denmark just before the COVID-19 epidemic. While in lockdown, inspired by curator Majken Overgaard's comment about admiring when artists re-visited old work, I returned to the concept of Book Reads You in a project called Zoom Reads You, supported by Nordic Summer University. As physical contact was limited or not permitted during this time, there was a surge in the use of videotelephony for communication. I was interested in the limitations of this form of communication and the concept of "Zoom fatigue" as well as how this was related to constantly being viewed at an unusually close distance by others who were "reading" your face. To this end, my hypothesis regarding "Zoom fatigue" was that by using videotelephony, users' digitized faces, through their micro and macro expressions, were consistently telling a story and that (as in Book

Reads You) this was not a story the user always wanted to be telling. To address their concerns, Zoom users had to strive to be aware of and control their expressions (for example, presenting themselves as professional and alert even if they were secretly in pajama pants). At the same time, facial recognition technology and large language models (hereafter LLMs) were becoming popularized.

I created a Javascript and PHP-based program using FaceAPI.js (face-api.js, n.d.), a computer-vision-based facial expression recognition API with options to use the 68 Point Face Landmark Detection Model, and developed a browser-based installation that recorded audio and webcam footage of the viewer with the ability to tell an audible story based on their facial expressions. The facial expression identification used by FaceAPI.js was derived from Ekman and Friesen's theory of facial expression labeling, which identifies six expressions: anger, disgust, fear, happiness, sadness, and surprise (1971) although FaceAPI.js included the expression "neutral" as well. Fifteen respondents participated in 5-minute recorded sessions that were saved automatically to the server for later analysis. For transparency, participants were also invited to download their videos. The text output from the program was also recorded and saved for analysis. Following their interactions with the *Zoom Reads You* program, participants completed a short survey about their experience. The text for the story was generated using a model of GPT-2 fine-tuned on sentences that contain the word "you" that I collected from text written in the second person. These texts included works of Fiction, self-help texts, astrology readings, meditation guides, as well as outputs of the GPT-2 model fine-tuned on these. Below is the outcome of one of the sessions:

"This is just a little brushing up in case you don't recognize yourself. You can't make out the answer with certainty. Maybe you are tired. Go ahead and stretch out your legs. In the pause you become aware of the rhythm of your breath the beating of your heart and you feel for a moment all the subtle motions that provide for your presence. What if it's about time your number came up? You're aware that something is bothering you but are somewhat skeptical that it will be noticed. You don't want it to be in fact. It is something to see alright. You had it just a minute ago. You manage to control your own image quite a bit. You have already seen how this might go and are prepared for what might happen. You fall silent as you wait for the next words your breath and gaze steady. Try as you might you cannot concentrate and find your mind wandering. Hmm perhaps you could have coordinated it a bit better but you have at least expressed the main idea. It's not that you can't see the past in your heart. Does that make you sad? Is this enough to say you would like to live several lives simultaneously? The existential crisis you face has nothing to do with you. Perhaps you haven't yet but you will in time. A light leaps to your face. You make steady progress toward your burning ambition. You feel a little bit like you were in a dream right now. To what extent do you find this agreeable? You could at least give a little thought to what you would describe to yourself as the feeling of being in a particular position. There are plenty younger than you or less young who live in the expectation of extraordinary experiences: from books from people from journeys from events from what tomorrow has in store. But there are things beyond your control. You identify with the texture of a person's stomach. A knowing look passes your face as you smile ear to ear. All's adequate after you've checked over it once or twice. You know something is not quite right. You're coming to grips with a new feeling. Are you being watched? You could help make the day worse. You're already exhausted today. You wish people would ask you your opinion more. You have never been able to convincingly demonstrate how much of your life have you spent sitting in this exact position. You're surprised to hear this. You try to think about it all logically. Ah you are participating... This is the conclusion you have reached in your personal life and also in general matters." -Output from a participant's session with *Zoom Reads You* (Jones, 2020)

Again, I found flaws in the limitations of the installation to "read" the person in an intriguing way. One of the most noticeable of these was that if a participant was initially identified as happy, the model would respond with an upbeat statement, which would make the participant happier, resulting in another upbeat statement. This would continue throughout their session. These participants stated in their survey that they enjoyed their interaction with the *Zoom Reads You* program. If, on the other hand, the participant initially appeared sad, the statements would continue to follow with depressing statements, making the participant sadder. These participants noted they did not enjoy the experience in their survey. The same pattern held true for the other expressions identified by the model. The model also "read" initial expressions incorrectly (for example, I apparently always look sad even when I have a neutral expression). Some participants tried to talk back to the program to correct or argue with it, but this did not change the output, as the program could only interpret expressions and not respond to auditory input. Ultimately, the boundaries of the language model combined with the limitations of the facial expression recognition program prevented it from providing a richer narrative experience, resembling the same problems presented in Book Reads You of the program's inability to understand greater context and situated experiences.

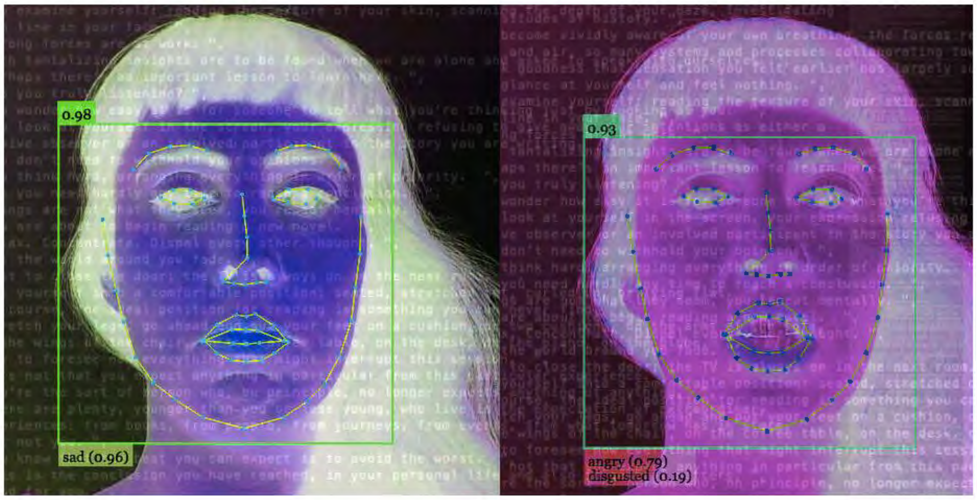


Figure 4. Zoom Reads You (2020) Mirabelle Jones supported by a grant from Nordic Summer University.

My experiences with these projects and others that questioned technologies, bodies, and the ability of technologies to tell people's stories ultimately led me to pursue my PhD in Computer Science at the University of Copenhagen in the section of Human-Centered Computing. In my work, I hoped to address questions as to how people formed relations to responsive AI systems in consideration of their identities (their stories) and how a plurality of individuals was considered (or not considered) by the developers of these systems. Based on my background as an electronic artist, software developer, and writer, I decided to use a speculative participatory art-based research approach to investigate these questions and concerns in HCI.

Bibliography for the Preface

Ekman P and Friesen WV (1971) Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology* 17(2). US: American Psychological Association: 124–129.

face-api.js (n.d.). Available at: https://justadudewhohacks.github.io/face-api.js/docs/index.html (accessed 8 July 2024).

Jones M (2020) *Zoom Reads You*. Available at: https://www.mirabellejones.com/zoom-reads-you/ (accessed 9 July 2024).

RACTER (1984) *The Policeman's Beard Is Half-Constructed*. 1st ed. Grand Central Publications.

Searle J (1984) *Minds, Brains, and Programs.* University of Michigan: British Broadcasting Corporation.

Table of Contents

PART ONE: RESEARCH DEVELOPMENT

INTRODUCTION 1

OBJECTIVES 10

- Research Questions 10
- Speculative Participatory Art-Based Research Method (SPABR) 11

METHODS 13

- Motivation for Art-Based Research 13
- Art-Based Research as Knowledge Creation 16
- Art-Based Research in HCI 17
- The Importance of Participation 18
- Speculative Participatory Art-Based Research Method (SPAR) 19
  - Speculative Art: Extending what is to what if? 20
  - Participation: extending art-based research outward 24
  - Making sense of how people make sense of responsive AI 28

PART TWO: SUMMARY OF PAPERS AND ANNOTATED PORTFOLIOS

SUMMARY OF PAPER I 33

AI is not your collaborator 35

Opportunities to push back against AI are essential 37

ANNOTATED PORTFOLIO I 39

Introduction 41

Insight: Unacceptable actions rejected or reinterpreted as acceptable actions 42

Insight: Repetition of programmatic logic in actions, materials, movements 45

Insight: Impossible actions recontextualized as possible actions 49

Insight: Vague instructions reframed as tangible outcomes 52

SUMMARY OF PAPER II 55

ANNOTATED PORTFOLIO I 58

Insight: Algorithms flatten data structuring perspectives 59

Insight: looking at you looking at me looking at me 62

Insight: Beyond devices & interfaces, inputs & outputs 68

SUMMARY OF PAPER III 71

ANNOTATED PORTFOLIO III. 75

Insight: The personalized bot as a fine-tuned mimic 76

Insight: Multisensory experiences expose questions of form vs. body 79

CONCLUSION **84**

REFERENCES **87**

PART THREE: RESEARCH PAPERS

PAPER I. **104**

Jones M, Neumayer C and Shklovski I (2023) "Embodying the Algorithm: Exploring Relationships with Large Language Models Through Artistic Performance." *In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 19 April 2023, pp. 1–24. CHI '23. Association for Computing Machinery. https://doi.org/10.1145/3544548.3580885

PAPER II. **129**

Jones M, Neumayer C, and Shklovski I (Under Review) "'It's Time We Talked': interactions with the self through personalized deepfakes." Submitted to New Media and Society. July 18, 2024.

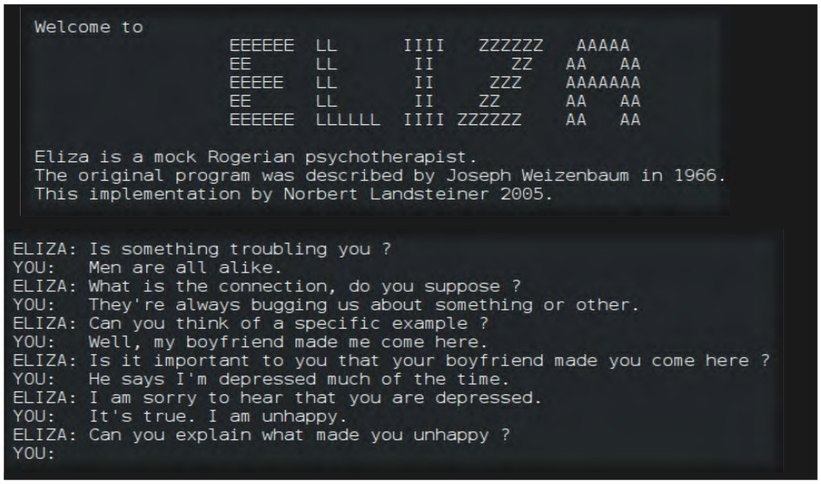
PAPER III. **163**

Jones M, Griffieon N, Neumayer C, Shklovski I (To Submit) "Artificial Intimacy: Exploring normativity and personalization through fine-tuning LLM chatbots." Targeted for submission ACM Journal of Responsible Computing.

Introduction

As artificial intelligence (hereafter AI) continues to become further entangled (Frauenberger, 2020) in people's everyday lives, information indicative of human-machine interactions is captured and aggregated to produce expanding assemblages of categorized (Bowker and Star, 1999), codified, and stored data (Scheuerman et al., 2023). The feedback loop of training and retraining through fine-tuning algorithms based on a continuous influx of accumulated data, that from the algorithm's viewpoint is all that is "us," results in outputs that force us to reflect and react in ways where we struggle to locate, make sense of, and gain control of these digital representations of the self. We are at times curious, cautious, or just creeped out (Shklovski and Grönvall, 2020) about the data that is both visible and invisible (Neumayer et al., 2021) to us about us, questioning how Target's online shopping algorithm can tell if you're pregnant (Hill, 2012) or how to rectify accusations of fraud because a government-supported algorithm flagged you as high-risk for being a single mother on welfare (Burgess et al., 2023). Even if we are aware of the numerous instances of algorithms destroying people's lives at a massive scale (O'Neil, 2016) or being biased in regard to race (Benjamin, 2019), gender (Ada Ada Ada, 2021; Buolamwini, 2022; Costanza-Chock, 2018; D'ignazio and Klein, 2020), ability (Forlano, 2023), sexuality (Gershgorn, 2017), and other categorizations of minoritized identities, we continue to interact with the products of algorithms at the cost of our data. With the rise of big data, fine-tuning through personalization, and generative AI, we are seeing an increase in the products of algorithms that respond to us in ways that simulate human interactions to the point where even developers can be duped into proclaiming the sentience of a large language model (Tiku, 2022). However, no matter how big the data is or how much of it is derived from human interactions, these systems fall short of producing human-machine interactions that are the same as human-human interactions because human interaction is fundamentally situated (Suchman, 1987) and context-based (Weizenbaum, 1967).

Nonetheless, from the first chatbot ELIZA created by Joseph Weizenbaum in 1966 to present-day machine learning technologies that use generative AI and are responsive such as ChatGPT, it is evident that people not only make use of these systems pragmatically as tools but also form relations to these systems if they are convincingly capable of mimicking a human. Even if ELIZA was not designed for the purpose of forming such relations but rather to study human-machine communication systems (Shrager, 2024; Weizenbaum, 1976), users formed emotional relations with ELIZA due to their tendency to attribute human traits such as meaning-making and empathy to the program (Suchman, 1987; Weizenbaum, 1976). This personification was due to their trust in ELIZA as believably human, a phenomenon that has become known as the "ELIZA effect."



```
Welcome to
                  EEEEEE  LL      IIII   ZZZZZZ   AAAAA
                  EE      LL       II       ZZ   AA   AA
                  EEEEE   LL       II     ZZZ    AAAAAAA
                  EE      LL       II    ZZ      AA   AA
                  EEEEEE  LLLLLL  IIII ZZZZZZ    AA   AA

Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.
```

```
ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:
```

Figure 4. Screenshot of a conversation with ELIZA from https://en.wikipedia.org/wiki/ELIZA#/media/File:ELIZA_conversation.png (public domain)

Opinions differ among researchers when it comes to addressing the human tendency to react to and interact with responsive programs ways that present attachment or personification. Shum et al. argue that constructing chatbots that are capable of forming emotional and empathic connections to users is essential in order to design for social good (Shum et al., 2018), but the authenticity of these connections requires semantic understanding gleaned from experiences in the real world and society, something that Weizenbaum argued ELIZA could only give the illusion of (1966) and several other scholars have echoed since (Bender and Koller, 2020). Some researchers encourage the exploitation of personification as an approach to design (Janson, 2023), while others stress concerns about privacy regarding how data is stored and utilized (Awad and Krishnan, 2006; Cox et al., 2023). But even with present-day systems trained on massive data sets, the impossibility of forming semantic understanding when it comes to human-computer interactions is emphasized as dependent upon the epistemic boundaries of the machine, even for programs constructed to mimic human cognitive functions such as generative neural networks. For example, Bender and Koller (2020) contend that natural language understanding (NLU) and natural language processing (NLP) programs (even those built on big data) are only capable of responses based on form, not semantics. While some have argued that NLU is attainable with more training data (i.e. bigger big data), Bender et. al stress the limitations and harms of constructing large language models based on ever-scaling amounts of data, including environmental concerns as well as the inability for these systems to live up to promises of openness and diversity, avoid biases and toxic language, but they also stress the limitations of such systems to serve as anything other than at best “stochastic parrots” (Bender et al., 2021), a term which bears a strong resemblance to Weizenbaum’s reminder that ELIZA was merely a “translating processor” regardless of the “aura of magic” the program may have impressed upon others (Weizenbaum, 1976).

Chatbots are just one example of the output of responsive algorithmic systems that people form relations to, but there are several other responsive technologies, including those that are visual, auditory, and tactile, that are the products of algorithms that raise similar concerns. One example is VR avatars captured from real-world data. In the South Korean documentary *Meeting You* ([VR Human Documentary] Mother meets her deceased daughter through VR technology, 2020), Jang Ji-sung is “reunited” with her deceased daughter Na-yeon as a VR avatar. The avatar was created using 3D scans of a child with a similar appearance to Na-yeon, basing this resemblance on photos and videos. In the video, Jang Ji-sung bursts into tears while simultaneously attempting (and failing) to embrace her daughter’s synthetic body. She states, “I really want to touch you just once,” as her hands go through her daughter’s avatar, acknowledging the lack of embodiment that identifies the representation of Na-yeon as the output of a program. In a follow-up interview about her experience, Jang Ji-sung acknowledges the fictional nature of the interaction in that the representation of her daughter is based on the data collected when her daughter was alive (age 7) and that at the time of the VR interaction, Na-yeon would have been eleven years old (Virtual reality ‘reunites’ mother with dead daughter in South Korean doc, 2020). Despite these acknowledgments of the synthetic and counterfactual, Jang-Ji has an emotional reaction to the avatar, stating that she was still “happy to see her that way.” The director also expressed concerns about the resemblance between the actual and simulated Na-yeon and how this might impact the mother’s experience. He stated, "I was worried how the mother would react when she sees the digitized Na-yeon. No matter how hard we tried to make the character similar, she still can tell the difference. But she said she was happy to see even the slight reflection of Na-yeon" (koreatimes, 2020).

Meeting You has been widely discussed as an example of an emotionally manipulative use case of interactive technology (Hadan et al., 2024) based on the lives and data (in this case photos and videos of Na-yeon) of real people. While Weizenbaum had no intention of producing a specific reaction, emotional or otherwise, from participants using ELIZA, *Meeting You* was intended to produce an intense emotional response and succeeded in doing so despite the noticeable flaws in the technology and resemblance of the avatar in comparison to the deceased child it was based on. On the other hand, ELIZA’s only “personality” was the role of a fictional Rogerian therapist, and ELIZA still elicited emotional responses from participants. The differences in social-technical implications of these interactions are worth considering. How does the choice of output (visual, text-based, auditory), the ability to respond and receive a response, and the fact that the program is based on real people's data change the way participants make sense of it and ultimately indicate the relations the participant might form with it?

In my research, I argue that it is crucial to address these concerns in order to understand how to develop and implement technologies from a position of care (Puig de la Bellacasa, 2017) and concern for how AI systems can replicate or exacerbate biases and further inequalities (Benjamin, 2019; D’ignazio and Klein, 2020; Eubanks, 2017). I am interested in how people make sense of and relate to responsive AI systems, including those that attempt to direct, mimic, or create alternate versions of them. Researchers and developers alike have engaged with the topic of the AI clone, also known as the digital twin, digital actor, digital doppelgänger, mimetic model, or synthetic self. (McIlroy-Young et al., 2022) speculate on the ethical implications of creating such models, identifying four distinct roles at play: the target (the person being imitated), the creator (developer), the operator (user of the model), and the interactor (the person who engages the model in some form of interaction). Most relevant to this thesis is their discussion of the potential purposes as well as ethical considerations for when someone is both the developer and deployer of a model based on their own data (McIlroy-Young et al., 2022). The ethical concerns they speculate on are related to disclosure (transparency), value (fidelity), and impact (including the potential for the model to behave in ways that would negatively impact the person’s reputation or relationships with others). However, their research, common to much research about AI clones, does not consider how one might make sense of and form relations to their own mimetic model and the ethical concerns that follow.

There are several similar studies that use qualified and/or quantified methods to address the potentialities and limitations of responsive AI models that mimic human-to-human interactions. Many of these studies are interested in improving transactional exchanges with responsive AI systems, such as creating chatbots in the fields of healthcare, education, customer service, and companionship. It is not surprising that the majority of research in responsive AI is geared towards producing better transactional experiences, as this allows companies and organizations to improve their reputations by appearing future-forward, minimize tasks (Park and Kim, 2022; Sako, 2020), and ultimately increase their profit margins (Kant, 2020). Similarly, the use of synthetic data to reduce the complications of gathering participant data in studies and trials has transactional motivations (Agnew et al., 2024) and can result in an increase in false positives that raise ethical concerns related to misrepresentation and complications for governance (Whitney and Norman, 2024). Much of the research concerning training models using real people’s data, or creating synthetic data to mimic this, fails to address interactions with AI systems that are not motivated by producing improved transactional AI or addressing the uncertainties surrounding AI systems, let alone viewing these uncertainties as fruitful resources for informing better HCI practices, design, and development (Benford et al., 2006; Gaver et al., 2003; Shklovski and Némethy, 2023).

This thesis presents three papers and an annotated portfolio (Gaver and Bowers, 2012) leveraging participatory art-based research and speculative methods to bridge the gap in understanding the implications of how people make sense of and form relations to responsive

AI systems that are built on a corpus of human-derived data. Through three articles that use grounded theory and mixed qualitative methods, I present findings from speculative participatory art projects created between the years of 2021 and 2023 that investigate how socio-technical experiences become personal through people's situated (including intellectual, multisensory, and emotional reactions) with responsive AI systems that are built on human data. I also explore in-situ and reflective sense-making processes that result from these interactions, provoking considerations regarding the historicity, present, and future design of these systems. In my research, I consider and invite others to consider the embedded politics and values enmeshed in responsive AI, in three explorations of LLMs, deepfakes, and personalized and fine-tuned models, through speculative participatory art.

Interactive and participatory art can provoke previously unasked and unanswered questions as well as challenge traditional approaches in the sciences toward knowledge formation (Jacobs et al., 2015). Both the benefits and complexities of incorporating interactive art as a research method in HCI have been discussed for decades. In 2003, Höök et al. created the Installation Machine to demonstrate the challenges and benefits of adapting traditional "gold star" HCI approaches, such as user testing studies, towards the evaluation of interactive artworks while acknowledging that establishing methods of doing so for HCI researchers and artists are often as different as the motivations for producing the work itself. Even so, they conclude that while evaluations of art as research should not be "forced to follow HCI principles," (Höök et al., 2003) there is much that can be gleaned through the co-mingling of the fields. Over twenty years later, the role and potential methods of incorporating art in HCI research are still being discussed and debated, often by exploring individual methods as approaches to expanding HCI. In this dissertation, I contribute my own method, speculative participatory art-based research, SPAR, as a productive approach to interrogating the relations people form with responsive AI systems.

As a method, speculative participatory art-based research's "findings" can be considered the conceptual basis for the work, the process of producing the work, the artwork itself, and the reactions of the artist and participants to the aforementioned. With the rise of new technologies that can serve as avenues for artistic exploration, including 3D printing, XR, AI (generative and otherwise), simplistic DIY prototyping materials such as microcontrollers and kits, open access programming interfaces (including those made by artists for artists), open-source programs and code, and finally the growing acceptance of interdisciplinary work among scholars, participatory art is more frequently used as a method to raise valuable questions through unique research approaches that can make important contributions to HCI. Participatory art encourages plurality and diversity in responses, embraces uncertainties and ambiguity (Benford et al., 2013), and recognizes messiness (Mackney and Young, 2022), discomfort (Benford et al., 2012), disruption and failure (Andersen et al., 2018), as valuable sites of knowledge formation. Participatory artists are not afraid to push questions concerning how people relate to technologies to their limits by stressing sense-making through situated and embodied interactions (Benford et al., 2021). As has long been discussed in participatory design, the participatory element is key to observing how people respond to the work and how it feels to those around them. Researchers using participatory art embrace inconsistencies, vagueness, and confusion in their findings as elements of intrigue and knowledge formation, while many deterministic practices in HCI try to avoid these. The same potential for beneficial messiness through approaching unasked and unanswered questions is true of speculative methods in combination with participatory art, daring to ask, "what if?" while participatory art also approaches the question of "how?" in a way that acknowledges the situatedness of responses. These combined approaches are appropriate for addressing the epistemic gap of machinic and algorithmic processes that attempt to mimic human beings but are unable to understand the important messiness and ambiguities of human experiences (Birhane, 2021b).

As the findings of speculative and participatory art approaches can be difficult to translate into research outputs in more conventional HCI formats such as articles and manuscripts due to their emphasis on respecting a plurality of situated truths gleaned from tacit knowledge and aesthetic interactions, the annotated portfolio serves as a way of both creating and disseminating knowledge. While publications and other textual forms of research outputs stress what Gaver and Bowers identify as the "indexical character" of these research outputs, the annotated portfolio focuses on text as insights in a conceptual space where artwork and description are "mutually reliant to create meaning" (Gaver and Bowers, 2012). Rather than focus on developing a specific methodological approach to an object based on materiality or pragmatic use as a means of cementing a research contribution (an approach that can place deterministic limitations on what could be conceptually abstracted from the object and people's experiences with the object), the annotated portfolio is a way of valuing the "particularity and multidimensionality of design work while meeting many of the demands of generalizable theory," (Gaver and Bowers, 2012). The annotated portfolio works well with qualitative methods such as grounded theory, as is demonstrated in this dissertation, as there is an emphasis on staying curious (Muller, 2014: 25) in order to discover variables as markers in a space of liminal conceptual departure rather than trying to control data with the result of "quantifying phenoma" (Corbin and Strauss, 2015: 853) according to a specific theory. The annotated portfolio emphasizes insights that highlight differences and similarities amid objects constituent to a body of work (Gaver and Bowers, 2012) that can help serve in comparison to other works as well as bodies of work. As (Gaver and Bowers, 2012; Löwgren, 2013) highlight, portfolio construction is already one of the artists' most predominant forms of research dissemination. Löwgren suggests that the annotations are "constructs residing in the in-between space" (Löwgren, 2013) in that they may be shared to different degrees of abstraction as annotations among different objects vary in presence or absence within that body of work for specific objects. Additionally, the portfolio itself may afford these abstractions.

Using SPAR as a method that evolved over the course of my PhD and the annotated portfolio as a crucial approach to demonstrating resulting insights, I formulate conclusions grounded in relational ethical concerns (Birhane, 2021a) from the standpoint of "staying with the trouble" (Haraway, 2016) to address the hazards of normativity of responsive AI systems, including personalized models. I focus on how AI that uses human data, such as large language models and deep fakes, is engaged from a situated perspective (Haraway, 1988). Using SPAR, I explore the gap in the authenticity of visual and textual generative representations that mimic human participants by using models trained on human data, including the ways these systems fail to "know" us despite presenting reasonable or convincing digital mirrors. Stressing the importance of intellectual, aesthetic, and embodied knowledge, I explore methods of sense-making used by participants engaged in a speculative participatory art practice involving responsive AI and what relations result from these as a means of diverting from a hyper-focus on technical fixes to techno-social issues. Furthermore, I consider the ethical implications of both these sense-making processes and relations in terms of state-of-the-art research in HCI regarding generative models that encourage personification.

Objectives

This dissertation contributes to ongoing research towards considering entanglement in HCI (Frauenberger, 2020) through the proposal of the method SPAR (Speculative Participatory Art-Based Research) developed over the course of three projects and papers herein discussed and presented through discussion, research outputs, and an annotated portfolio. Embracing non-linear (Kang et al., 2022) and disruptive practices (Andersen et al., 2018) and the messiness of art-based research (Mackney and Young, 2022), the research questions are addressed throughout the dissertation.

Research Questions

Through three research papers and an annotated portfolio this dissertation addresses the following research questions:

RQ1.1 How do humans make sense of responsive AI systems?

RQ2. What are the results of using speculative participatory art to explore human relations to responsive AI systems?

To address these questions, I created three artworks that served as the basis for SPAR projects to explore specific aspects of these:

I. ***What happens to the body when people use text-based responsive AI systems such as LLMs to direct human behavior?*** *(Addressed in paper I.)*
II. ***What happens when responsive AI systems become personalized to include personal visual data such as deepfakes?*** *(Addressed in paper II.)*
III. ***How do people respond to foundation models vs. personalized models trained on their social media data?*** *(Addressed in paper III.)*

SPAR Research Method Contribution as Demonstrated Through Projects

This thesis presents three papers that explore the relations people form with AI systems using SPAR while ultimately questioning how these relations are considered (or not considered) by the designers of these systems from a situated perspective that challenges normative design. It also considers the resulting socio-political implications of how these systems exploit sense-making processes and relations. In three participatory art-based projects using speculative methods, I explore: 1) the implications of large language models directing the body through endurance performance art, 2) making sense of deep fakes using science fiction, and 3) personalized chatbot systems using fine-tuned large language models based on personal social media data.

Project One: *Embodying the Algorithm.* This paper explores what it is like for large language models to direct human behavior, ultimately revealing two findings: 1) because AI systems are incapable of understanding the human body, it is necessary for humans to have the ability to push back against the directives of AI systems. 2) People tend to personify these systems and want to form relations with them, but because language models cannot respond in kind, relations can only be formed to them.

Project Two: *It's Time We Talked.* This paper demonstrates what it is like for people to encounter their own deepfake double using an approach inspired by science fiction where participants meet a speculative version of themselves from another timeline. This project revealed that even after an educational workshop about deepfake creation and a discussion of ethical concerns held at Catch in Helsingor, Denmark, participants experienced a two-step process of sense-making that included both self-reflection as well as external validation. Our analysis of the interview data revealed tensions regarding the imagined perspective of how machines see people in contrast to how people see themselves, as well as tensions between autonomy and control over these systems. We conclude that the SPAR process allowed us to “defang” deepfakes to deconstruct these tensions in order for participants to pursue their own situated relations to deepfake.

Project Three: *Artificial Intimacy.* Exploring normativity and personalization through fine-tuning LLM chatbots. Through participatory art-based research and a series of four workshops, this paper explores the implications of personalization and normativity in chatbots that are fine-tuned on people's social media data vs. base models (in this case GPT-3's Davinci model). As suggestions have been made that personalizing through fine-tuning large language models can improve users' willingness to engage with them and result in value alignment, we questioned what the experience would be like for people to engage with fine-tuned models based on social media data through a process of personalization. We find that while there may be a preference for participants to prefer chatting with fine-tuned personalized models because they seem more "human," there are simultaneously opportunities for uncanniness as well as intimacy that emerge from engaging with chatbots that mimic the self.

Overall, this thesis contributes the method SPAR as demonstrated by conducting research into how people engage with different responsive AI systems and speculates as to how this may further research as HCI incorporates interdisciplinary practices and methods to accommodate new approaches to knowledge formation that address entanglement (Frauenberger, 2019). It suggests how to challenge, through participatory involvement, the normative design of generative AI products that result from algorithms constructed on majoritive human data while also considering the impact of personalized, responsive AI to invite experiences that open opportunities for uncanny and intimate experiences. Lastly, and most importantly, this thesis contributes insights into how people relate to responsive algorithm-based technologies using qualitative research methods to analyze first-hand experiences in consideration of how these technologies connect or depart from our understanding of our own situated identities.

Methods

Motivation for Participatory Arts-Based Research

As an artist who has used performance and interactive art separately and in combination for over a decade, the relational aspect of my work has always been an important part of what I learn from my practice in interacting with people and technologies. However, it took time to reflect on my role as an educator and activist to really appreciate the research contribution of the participatory element of my body of AI artwork to HCI.

After moving to Denmark from the USA in 2019, I became increasingly interested in probing questions about the role of speculative technologies in the arts (and vice versa) and held several workshops and courses for organizations like Gray Area Foundation for the Arts, Catch: Center for the Art, Design, and Technology, Artsformation, Codame Art & Tech, and HackadayU. I also presented my findings from my participation in Nordic Summer University of the online installation Zoom Reads You (Jones, 2020), where I had the chance to engage with researchers and artists from diverse backgrounds and present my findings in the context of academic research. Observing how workshop participants, students, and researchers each brought unique questions and approaches to engaging with technologies, I thought I knew the limits of made me reflect on the importance of interdisciplinary perspectives to challenge socio-technological assumptions.

I also gained an increased interest in participatory art as a means of activism and protesting the injustices embedded in social norms by reflecting on my former role as founder of Art Against Assault. Art Against Assault was a USA-based organization that guided the production of art-based workshops and exhibitions for survivors of sexual assault and domestic violence. I operated Art Against Assault starting in 2011 as a way of raising discourse about these topics prior to the #metoo movement. One moment that resurfaced frequently was when I collaborated on an exhibition of student survivor work at the Rhode Island School of Design. One student had created a tapestry with missing stitches that hung in the middle of the gallery. She explained how the work expressed her feelings towards her experience of sexual violence and told me that her art instructor had critiqued her harshly for it, demanding to know why she wanted to “traumatize others” with her work. The instructor further violated the student by institutionally discrediting the value of an artwork born of a lived experience of violation. Experiences like these led me to consider how institutions discredit the value or research born of emotions, embodied experiences, and taboo or non-normative subjects.

The culmination of reflections on my previous role as an activist for voicing realities that were socially and institutionally silenced and making them valid and visible through artistic practice, coupled with witnessing how interdisciplinary approaches can enrich conceptual discourse in workshop settings, propelled me to fully grasp the importance of the participatory element in arts-based practices and further my drive to archive, disseminate, and legitimize this knowledge through academic research.

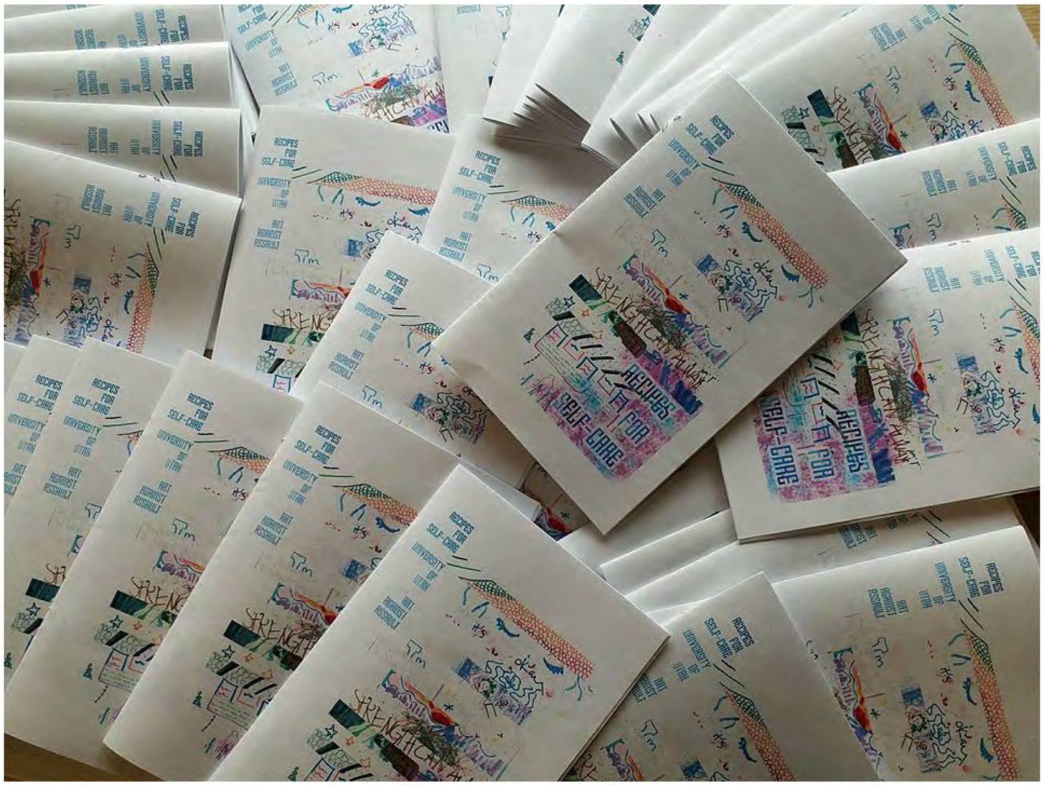


Figure 5: *Recipes for Self-Care* zine produced by students at the University of Utah in a workshop led by Mirabelle Jones for Art Against Assault (image from ArtAgainstAssault.com)

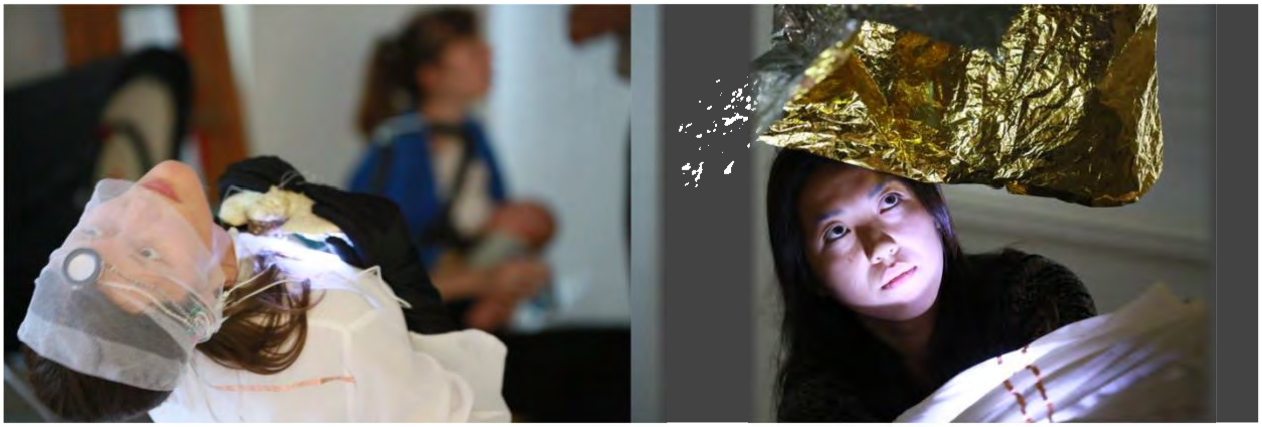


Figure 6: Participant work from the Soft Circuits Summer Camp taught by Mirabelle Jones at Catch: Center for Art, Design, and Technology Education. (Image from MirabelleJones.com) Photos by: Frej Schmedes of a performance by Linh Le and Ida.

I was also influenced by my previous role as a Senior Designer and Developer for Meow Wolf where I aided in building interactive artwork for warehouse-scale experiences. Although the artworks were often absurd and odd, they were expected to work in more-or-less predictable ways and endure interactions with thousands of visitors daily. Although I enjoyed making speculative interactive art every day for a living, the process of engaging with industrial expectations where experiences should function as smooth-sailing state machines left me longing for the opportunity to create interactive works that were less deterministic in their design motivations. I also learned through user testing that if you let people interact with an object (especially a responsive one), they will do all manner of unexpected things (a reality that excited me). When testing an interactive 3D-printed sculpture designed for capacitive touch response to see how people would touch or grasp it, one participant licked it, and another stuck their nose in it (from these experiences, I formulated the theory that if you want to truly generate randomness, use people, not math). While I was forced to see these experiences as undesirable or failures in my role as a designer and developer (perhaps similar to how an HCI researcher using deterministic methods might view unexpected or messy outcomes), I yearned to acknowledge the unpredictable nature of human interactions with technology, embrace the importance of the body, and push back against normative assumptions.

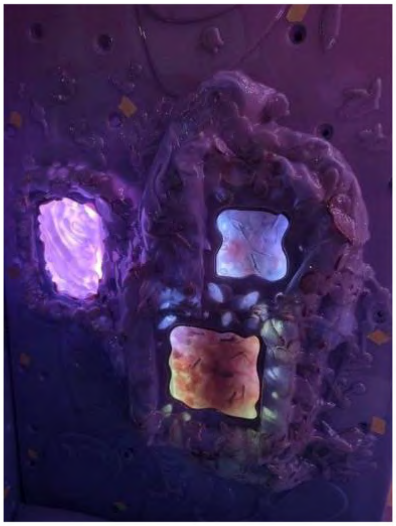


Figure 7. The purple sculpture to the left of the image is a 3D-printed interactive design with embedded LEDs that respond to capacitive touch. To find out how people might touch it, user testing was conducted in which participants put their noses in and licked the sculpture. Sculpture designed by MirabelleJones. Image from MirabelleJones.com

Art-Based Research as Knowledge Creation

Art-based research (hereafter ABR) is a qualitative inductive method grounded in artistic practice that highlights emotional, multisensory, and intellectual human experiences as not only valid but important sources of data. ABR constitutes a wealth of ever-expanding approaches that can include presenting data related to the conceptual evolution of a work, data about the work's construction, data about the work as a realized object, data about its exhibition, and data about responses to the work. While also employed by artists in more traditional studies in the arts such as painting, dance, and sculpture, ABR as a method grants a transdisciplinary group of scholars working with new technologies an opportunity to challenge institutional assumptions and evaluation methods, inviting unprecedented approaches to formulating research through conceptualizing, conducting, and capturing a wide range of artistic practices. At its core, ABR arose from a desire to challenge the epistemological underpinnings of traditional methodological approaches in academic institutions. Knowles and Cole define ABR as a demonstrative act of protest against scientific traditions of knowledge formation, stating, “to claim art and aesthetic ways of knowing as research is an act of rebellion against the monolithic ‘truth’ that science is supposed to entail,” (Knowles and Cole, 2022). At the same time, not everyone who uses ABR shares the standpoint that the process is, by nature, in contrast to or in opposition to more traditional forms of knowledge creation within the sciences. Because ABR emphasizes the value of cognitive, bodily, and affective knowledge, it has frequently been used in healthcare and education (often in combination with quantitative methods and more traditional research approaches), as well as by designers who value artistic processes as resources for understanding the felt experiences of others as materialized through artwork.

Art-Based Research in HCI

In HCI, art-based research is becoming better acknowledged as a form of knowledge creation, particularly as HCI continues to expand to incorporate multidisciplinary materials, practices, and approaches. While some of the same lingering questions arise about how art can be evaluated within HCI's established frameworks and systems, there are many scholars in HCI who see the advantages of re-framing research as not only something that is done (where the results are equivalent to the value of the research output), but that research is a process that is felt and thought about which can facilitate knowledge and sense-making (Jungnickel, 2018, 2020). Some of these scholars look to interpret established approaches within art-related or process-related fields as opportunities for incorporation into HCI's already accepted ways of knowing. (Frankjær and Dalsgaard, 2018) successfully demonstrate how Richard Sennett's tripartite deconstruction of the crafting process can be applied as a mode of inquiry in HCI using three artworks as examples. (Kang et al., 2018) similarly demonstrate how valuable insights found in improvisation studies can be incorporated into a collaborative art practice in HCI. This suggests that incorporating existing art-based research approaches into HCI invites new analytical opportunities for conceptualizing research. Other scholars demonstrate embracing non-linear processes (Kang et al., 2022) and non-deterministic practices (Andersen et al., 2018) that do not equate to an absence of process or plan (Dourish, 2001: 71; Suchman, 1987) or a lack of meaningful outcomes but provide the opportunity for more robust possibilities for these. Of course, research processes and strategies that embrace approaches such as these might result in a plurality of meanings that can be misconstrued as messy or inconclusive research (i.e. an insufficient contribution to the research community). In fact, these plural or even conflicting meanings can be, in many ways, more authentically representative of the complexities inherent to the research space

(Sengers and Gaver, 2006). These pluralities present opportunities for multiple points of departure toward further addressing conceptual research problems while also acknowledging conflicting truths about lived experiences that offer valuable data about human-machine interaction.

The Importance of Participation

Participation extends art-based research (ABR) outward, investigating how people hold and create their ideas through experiences taking as a point of departure the co-conception, creation, and/or dissemination of artwork while also encouraging participants to share their thoughts, sensory experiences, and/or feelings with researchers and each other (all considered valuable sources of knowledge). Participation can take the form of workshops, hackathons, performances, exhibitions, demonstrations, happenings, flashmobs, online collaborations, as well as any other artistic processes in which many are involved to participate. (Holmer et al., 2015) identifies three participation methods: spectacle, making, and inquiry, which the authors state may serve as points of departure for consideration of participation approaches in HCI. In my research, I am interested in how all three of these can be considered as processes in a methodological departure from participatory art-based research, SPAR (Speculative Participatory Art-Based Research) to explore relations and meaning making with responsive AI systems. As in ABR, in participatory art-based research (PABR), value is included in a plethora of experiences, including those traditionally seen as unimportant, undesirable, or to be avoided in research methods, such as those with defined deterministic underpinnings and/or utilitarian research goals. As such, there is an element of the participatory that is also political. There are debates in several areas of research about what constitutes participatory art, how participatory art reinforces, illuminates, or dismantles systems of power by creating hierarchies and establishing systems of exploitation under the guise of “co-collaboration” with the goal of improving people’s lives. In *Artificial Hells*, Claire Bishop (Bishop, 2023) takes as a point of departure from Rancière’s aesthetic dissensus to call for participatory art that is critical: disrupting systems of power and creating tensions that highlight lived inequalities, imbalances, and controls. PABR provides a resource for exploring interactions of this nature where ethics and the safety of the participant must be considered. In HCI, (Benford et al., 2012) discuss the aspects, approaches, merits, complexities, and ethical concerns of designing uncomfortable cultural interactions using artworks and performances as examples. They identify entertainment, enlightenment, and sociality as benefits of engaging with uncomfortable interactions while acknowledging that the discomfort must be contextually positioned as "a moment on a journey" rather than an enduring experience. Participation also allows for the acceptance of multisensory, emotional, and intellectual data that differ meaningfully among participants or participant groups. In regards to multisensory experiences, Kia Höök's explorations of somaesthetics (Höök, 2018) draw on Shusterman's proposal of the approach as "the critical, meliorative study of the experience and use of one's body as a locus of sensory-aesthetic appreciation (aisthesis) and creative self-fashioning… devoted to the knowledge, discourses, practices, and bodily disciplines that structure such somatic care or can improve it" (Shusterman, 1999: 302). Höök not only reinserts concerns of bodies into discussions of technological design spaces (concerns that are frequently absent in discussions of technologies that are assumed immaterial such as AI and ML as addressed in paper I) but rejects the Cartesian dualist notion that embodiment is something that needs to be separately addressed in design, for "As there is never any chance for us to be disembodied, it does not make sense to speak about embodiment," (Höök, 2018: 48).

Speculative Participatory Art-Based Research (SPAR) Method

SPAR has evolved over the course of the making and analysis of the three art projects included as the subject of the research outputs of this dissertation. With a foundation in participatory art-based research, it is a method that can be used to explore socio-technical relations by identifying sense-making processes of personal encounters with artistic objects incorporating technology. As I demonstrate, these personal encounters can shift in how much the researcher scaffolds the participant's experience with interpreting, viewing, or programming the technology, as well as what knowledge participants already possess of the technologies and techniques at hand. Drawing inspiration from participatory art-based research's attention to multisensory, affective, and intellectual participant responses as valuable sources of data, participation is used to probe previously understudied areas of interactions within the context of HCI but can also be extended toward other domains to explore techno-social limitations and boundaries.

The method begins with a speculative approach to art-based research, which includes inspired speculation, theoretical and concrete experimentation, materialization (making), and exhibition of one or many art objects. Next, there is a phase of participation that can involve different forms of discourse and making, where participants are invited to use the artwork(s) as a scaffolding for creating or co-creating a personalized artistic work with a technology to engage with. This is followed by a reflective stage in which participants are questioned about their experience using mixed qualitative methods (in the projects in this dissertation, I use semi-structured interviews, field notes, and questionnaires). In the final research stage, the researcher takes up an inductive approach to identify sense-making processes and finally reveal relations participants formed with the technology. In this way, the researcher is able to draw conclusions about the limitations as well as benefits of the technology in terms of discursive claims as well as state of the art and consider implications for researchers and designers using these technologies.

Speculative Art: Extending what is towards what if?

My research begins by asking open-ended questions about AI invited by exploring works of art and discourse surrounding responsive AI technologies. I explore both contemporary and historical artistic works such as literature (e.g. science fiction in paper II and III), performance (e.g. paper I), interactive art (e.g. papers I, II and III), XR art (paper I), video (e.g. paper I, III), music (e.g. paper I), and sculpture (e.g. paper I, II, III), photography, fine art, and other forms of art that invite questions about the incorporation of emerging technologies. Grounding the research in artistic exploration puts the research in dialogue with artistic practices that are already responding to technology not only as an aesthetic material but as a point of socio-technological intrigue and speculation regarding human-machine relations. In this stage, I also investigate discourses surrounding technologies in research and the media to consider points of intrigue and contention in dialogues and debates at large to identify gaps and further speculate on possible situations that can be created through participatory art. Finally, I consider how the creation of the object might build upon my own body of conceptual and theoretical work, using my existing knowledge and skill not only as a researcher but as a multi-faceted human with various interests, talents, curiosities, and approaches to thinking and feeling. This initial speculative stage of research also provides a platform for ideation that guides me towards the creation of the artwork, transitioning from questions of “what if” to “how?”

As many interdisciplinary researchers have pointed out, the existing knowledge of the researcher is not limited solely to their own field of research. In my approach, I use my background as a storyteller and writer, educator, developer, and artist of many mediums (photography, performance, installation, book art and printmaking, video, creative writing, songwriting) as well as considering the goals and intentions of these works (activism, play, testing the boundaries of the body and mind, questioning identity and self-reflection, creating alternate worlds) to build my repertoire of tools that can be used to answer the question of how a speculative concept can become a participatory art object. I also must answer the question of “how” when focusing on what technology I wish to approach (be it a technology I am familiar with and am using in a new way or a technology I have never used previously). In this way, I learn new technologies and new applications for old technologies, including programming languages, hardware limitations, and capabilities. I also engage with existing research and artworks already involved in the discursive space of these technologies so I can eventually create the work and teach the process to others. My approach to combining artistic approaches with software development is to gain entry to impossible futures from an open, curious, and investigatory perspective rather than using the deterministic approaches more familiar to HCI practices. In opposition to designing an interaction where I want the “user” to feel X, I consider broadly what thoughts and feelings might come from interactions with an art object and what materials might enable these.

Going from immaterial speculation towards processes of making is an opportunity to experiment with new technologies and mediums to discover possibilities and limitations for learning, intrigue, discursive spaces, conceptual grounding, and feasibility. The inability to grasp or use a specific artistic medium or technology (even with the benefit of having access to a large repertoire of inspiring projects – many accompanied by code – on sites like GitHub and various maker and hacker blogs e.g. Hackaday, Hackster, Arduino, Adafruit, Instructables) sometimes results in a need to plan with others. As an artist who works with technology, I have learned the benefits of acknowledging when others are more knowledgeable about a skill, medium, or technology and often accept DWO (doing with others) over DIY (do-it-yourself) in the making phase through skill-sharing with other artists, developers, and creatives. While many artists and developers are encouraged to give the impression that their work is solely theirs, this is almost never the case. Artists and developers are always influenced by co-workers, fellow developers, and others within their social network and existing and previous bodies of work as sites of knowledge and inspiration. There is no such thing as solo work. By acknowledging this, the producer of the work can embrace the fact that the resulting work embodies a collection of knowledge. The object is ultimately a collaborative interdisciplinary work. Hence, participation exists throughout the process, from ideation/speculation to the process of making.

As many researchers and artists alike have addressed, making is a sense-making process or, as Jungnickel describes it, a process of "making things to make sense of things." By critically and creatively approaching new technologies and creating an art object representative of something that was once a concept dislocated from material, concepts can be extended outwards as an invitation into public discourse. The artwork, in this way, becomes participatory. Individual questions, observations, and speculations can become points of collaborative speculation while honoring the situatedness of different perspectives.

To this end, all materials of the object, including hardware and software, become sites for conceptual exploration. The sharing of code in this way also becomes participatory, as is highlighted in paper III, and excluding access to code results in an inability for people to gain transparency through experiential knowledge of software products, as is highlighted in papers I and II. Although code as art material has yet to be fully accepted in many computer science programs as science (resulting in the bulk of "creative code" research being ostracized to the humanities), the works of Winnie Soon, Geoff Cox, Kim Albrecht, Dan Shiffman and Lauren Lee McCarthy, and many others who extract elements of the liminal space between code and program to break the dichotomy of input/output, observing code as a site for techno-social exploration including the embedding of values, normativity, and power dynamics. In

Aesthetic Programming (2021), Soon and Cox discuss how existing approaches towards the development of technologies result in power imbalances not only in unrecognized issues pertaining to the social, such as questions of representation but how code is presented (or absent) in curricula. Similarly, Dan Shiffman and Lauren Lee McCarthy, through the creation of the programming languages Processing and P5.js as well as open-source educational materials and online platforms for code sharing, view code as not only a means to an end but a material with which to play with identity and how it is represented through technology. Kim Albrecht, who self-identifies as a "knowledge designer," probes the politics and boundaries of data visualization, exploring how layers of development processes represent decisions related to control and transparency. In *Laws of Zoom* (Albrecht, 2023), Albrecht breaks down the API as containers of potential performance, visualizing these as a network of possibilities and constraints.

In each of the papers presented, the participatory engagement as a process of making sense through making (Jungnickel, 2018) is expanded. In paper I, participants cannot access GPT-3 directly and, therefore, must participate with the technology through me as a mediator for co-writing sessions. In paper II, participants learn about the ways deepfakes are created and participate in contributing their video letters as destination material for a deepfake, but ultimately, I am the one who creates the deepfakes. In paper III, through hands-on workshops where participants are invited into the data creation process (compiling a dataset of their social media data) and development process (fine-tuning GPT-3 using their dataset), this exploration of code as a conceptual medium comes to the forefront (see paper III summary).

This does not mean that the visible materials of the finished object are remiss of their contribution to eliciting socio-technical concerns, as discussed in papers I and II. In paper I, the constraints of the program are made visible through the expression of the artist using performance depicted in videos captured using a webcam. As we highlight in paper I, and as is demonstrated in the annotated portfolio for paper I, performers use different approaches of sense-making to develop their final performances. Similarly, to Albrecht's work of *Laws of Zoom*, questions of constraints embedded in the epistemic boundaries of the program, are rendered visible through representations of embodiment. In papers I, II, and III, questions raised and addressed by Shiffman, Cox, Soon, and McCarthy about socio-technical power dynamics embedded in code, gradually evolve in the form of participation and the expansion of participation to include participants as programmers who see further behind the veil of technology into its components, becoming further integrated into the participatory process of making. Creating the initial artwork and understanding the materials (visible and invisible) as points of conceptual departure or frustration, allows me to consider the degree of agency in the creation of the artwork each participant might have. The initial artwork's creation and exhibition allow me to identify at what point in the creative process the participants are asked to contribute (intellectually, emotionally, somatically), to identify "sticky points" (potential points of stop-process confusion), as well as ethical considerations (what is ethical to ask of the participant in the process based on my own experience with creating the work? How might others who are of different identities have different concerns that I should address?)

Participation: extending art-based research outward

Participation begins with the object, but where the participation ends is reliant on the participants' actions and reflections following the workshop as well as the exhibition of the objects as extended sites for participation. In the three projects explored in this dissertation, participation began in each case with the exhibition of the object because the exhibition of the object is part of the dissemination of knowledge that can be used to generate a platform for participation. The exhibition of the work offers up a demonstration to an individual audience or community of a possibility of making the speculative into something real. By doing so, it demonstrates to the participant that the translation from speculation to real is imperfect including flaws, limitations, and glitches that expose the impossibility of a "perfect technology." This alleviates the tension of working with unfamiliar technologies and artistic practices as well as encourages participants to reconsider dialogues of technological utopianism as grounded in the real expressions of these technologies. The act of exhibition is also political. As Rancière states in The Politics of Aesthetics, "Artistic practices are 'ways of doing and making' that intervene in the general distribution of ways of doing and making as well as in the relationships they maintain to modes of being and forms of visibility," (Raciere, 2004). Be it placing code on open-source platforms or an art object in a gallery, I am not only making decisions of where to situate my work which is a political act (for example, a closed exhibition vs. an open exhibition). The site of location creates power imbalances in who has access and is encouraged to participate. I am also contributing to a social-technical process of how work of specific mediums and aesthetics are normatively expected and accepted in certain communities. Exhibition is also an act of vulnerability. By sharing my work, I am extending my own flaws and imperfections and opening myself up to the communities in which I have chosen to exhibit my work. This vulnerability is essential when it comes to participation which requires trust as I will discuss in the upcoming sections.

In papers I, II, and III, each exhibition takes up a different site of exhibition moving from self-created online gallery spaces (exhibition at a distance due to COVID-19 resulted in the platform aiperformance.space) in paper I, to the physical gallery space Catch in paper II, to hybrid online and institutional spaces (Zoom, Harvard, and the University of Copenhagen) in paper III. Selecting these sites for exhibition offered different sets of expectations about what participation would entail, as well as placed limitations and opened possibilities for what participation could entail. For example, in the paper I, because of the limitations of exhibition spaces provided by COVID-19, participation, as well as dissemination, had to take place at a distance and online mediated through various software platforms. Once participants accepted the invitation to participate in the project, I used Zoom sessions to go over the project individually to each performer. In these sessions, I offered them a chance to view the four works for *Embodying the Algorithm* which I had performed, explained my process of creating them, detailed the limitations of the project (based on practical concerns and ethical concerns), and presented the Google document participants could select a prompt from or offered participants the option of co-writing sessions with GPT-3 over Zoom. I also demonstrated how GPT-3 worked by screen-sharing the beta playground and showing the probabilities attributed to the next token. At this point, the participant took over the creative process. The artist created the work in their home studio using their own materials and according to their own interpretations as well as on their own time schedule. They recorded their performance(s) using a webcam and the video was sent to me through a file transfer service of their choice. Finally, they participated in semi-structured Zoom interviews concluding their participation.

In paper II, participation took place in a gallery setting, garnering those who felt comfortable in gallery settings and had access to that location. Unsurprisingly, all the participants had some interest in the arts (they were either artists or designers or had some experience creating artwork), and all lived in Scandinavia. The participation took the form of a workshop in which participants were shown the physical installation of my work in the exhibition space (allowing a more intimate encounter than viewing it online) and were able to interact physically (an experience that further extended my vulnerability to them and them to me). The installation served as an example of a possible outcome of a deepfake (me deepfaking myself), while the workshop offered a space for participants to be presented with information about deepfakes, ask questions, create, and share knowledge with each other as a closed group. As noted in paper II, the workshop helped establish trust, which was vital for engaging participants in a form of participation that required more vulnerability between them and I than in paper I, where participants created their own interpretations of the rules and ultimately decided how to perform them, creating the final artwork themselves. Because the participants would be sharing both the visual data (in the form of recorded videos) and personal histories (the letters written to their alternate selves) with me to use in a collaborative process, they were, in fact, trusting me with two very personal forms of data to use in the creation of a personal artwork. As such, I did not interrupt them during the letter-writing process or direct the videos beyond suggesting they look at the camera as much as possible. After the workshop concluded, I created the deepfake videos and emailed them to the participants. As in paper I, follow-up interviews over Zoom concluded the participation.

In paper III, several types of participation occurred in different phases, including during the creation and exhibition of the initial artwork and in the four workshops that followed. Below I describe the process of creating the initial work and the kind of participation that involves as well as the participation platform, requirements, and challenges for the workshops.

Initial Artwork: In the initial artwork, rather than using my own data as I had in the first two projects, I produced a video installation and sculptural installation featuring two friends of mine, Leslie Foster and Gorjeoux Moon. I made the decision to not use my own data for two reasons. First, I was already very familiar with the fine-tuning process and chatbots, including personalized models of myself. Secondly, because the purpose of the project was to explore what it might be like to produce non-normative chatbots based on non-normative people's data, I felt it would be useful to explore other people's first-time interactions with their chatbot self and what they (as fellow artists but also as people who had not experienced personalized chatbots) might find interesting about the process. I presented the project over Zoom to Foster and Moon because they live in different countries. In order to educate them about the project, I showed them the GPT-3 beta playground and how it functioned and discussed my intentions for the piece so they could gauge the participation required and whether the expected outcomes would align with their body of work as well as availability. To produce the work, I needed their social media data and asking for this was indeed asking for them to put a great deal of trust in me. As long-time friends and collaborators, they agreed to this and I agreed not to read it before they had the opportunity to delete any conversations they did not want included. They downloaded their Facebook Messenger data and sent it to me using file transfers. Without reviewing this, I converted the data to a CSV file, which I sent back to them so they could remove anything they did not wish to use in the training or have read by me. I then fine-tuned both models using command prompts according to the instructions provided by OpenAI at the time. For the next step of participation in the artwork, I invited Moon and Foster to, in separate sessions over Zoom, ask questions of their fine-tuned models in the GPT-3 beta playground. What I found most interesting about this stage of the collaboration is that it confirmed my suspicions that incorporating other artists instead of myself was far more interesting in that it invited the kinds of different outcomes in interactions that considering situated knowledge through participation might allow for. The fact that both were artists with their own bodies of work, which had their own conceptual underpinnings, meant they asked very different questions than most people would ask of chatbots. They were also not afraid to ask questions that were quite intimate: questions of race, gender, sexuality, trauma, addiction, and abusive relationships, as well as interests explored in their practice: the motivation for poetry, Indigenous conceptions of time, black futurity. The resulting video installation shows clips of Foster's profile asking questions they've selected to their fine-tuned model, then switches to Moon who does the same. The bots were later incorporated into an installation for a demonstration at NordiCHI 2022.

Workshops I – IV: Workshops I and II were held online and workshops III and IV were held in a hybrid online and in-person format at the University of Copenhagen. As my colleague and I noted, both formats presented separate sets of issues for participation, especially considering the level of participation was once again heightened by participants taking even more control over the process of creating the artwork by programming it themselves using their own personal data. In workshops I and II., which were held online over Zoom, we invited participation using an online sign-up form where we also collected demographic information. Workshop one was held in conjunction with metaLAB Harvard during my change of research environment. However, this workshop had to be held online due to COVID-19 restrictions. My colleague and I decided to allow anyone to participate in the workshop, and it was advertised as not requiring any prior knowledge of programming or experience with NLP. A link was included to the video artwork produced with Leslie and Moon which was later shown at the Harvard Art Museum for those who wished to see it in person.

Making sense of how people make sense of responsive AI

Investigations into sense-making in HCI spanning the past thirty years have diversified approaches towards understanding the relations people form with technologies in how information is retrieved (Russell et al., 1993), towards observing sense-making as a situated eco-socio-technical process that is indicative of global power imbalances that can be explored through participatory practices (Bødker, 2015). One example of this is Ames's ethnographic study of the XO laptop campaign One Laptop Per Child (Ames, 2019). Ames contrasts the promotional dialogue and hype surrounding the One Laptop Per Child campaign with how the actual insertion of the laptops into their targeted environments played out, with particular attention to how the campaign impacted the lives of those who encountered the technology. Promoters of the campaign marketed the constructionist idea that placing a laptop in the hands of children in the global south would increase programming interests among children who did not have access to the technology. As Ames observed through ethnographic fieldwork, the laptops became burdensome in many respects. From their material deficiencies to the fact that teachers who were also unfamiliar with them didn't know how to provide guidance on how to use them to the fact that most children didn't instantly and without support (i.e. magically) find programming a meaningful interaction, the assumption that the introduction of the technology into an environment would result in meaningful relations was stopped short by a lack of infrastructure for sensemaking of the devices that would promote them as programming tools. As Ames points out, promoting the imaginaries of software development to potential investors as an extension of the myth of the natural programming genius (the imaginary of "the technically precocious boy" based on Western ideals of how great programmers come to be) was the primary focus of the creators of the campaign. The campaign promoters' interests did not stem from understanding the ways in which children in the global south (a large portion of the globe that includes an abundance of different communities) might make sense of laptops as meaningful technologies that could form a set of relations based on their situated encounters and, thus, they did not consider the research and infrastructure needed to do so.

Imaginaries like those that propelled the One Laptop per Child campaign fuel the hype of AI today while distracting from AI's realistic capabilities and limitations, as addressed by (Campolo and Crawford, 2020) in their discussion of "enchanted determinism." As described by Campolo and Crawford, enchanted determinism is a discursive approach that relies on supporting myths and imaginaries of 1) superhuman and 2) unexplainable AI to avoid responsibility towards addressing the realities of AI. These unaddressed realities include issues of transparency and explainability (Pasquale, 2016), environmental impact (Kirkpatrick, 2023), the exploitation of human labor (Sarkar, 2023), exacerbation of social biases (Hajigholam Saryazdi, 2024), worrisome military applications (Rivera et al., 2024), and capitalistic global power imbalances (Ricaurte, 2023), among an abundance of other hidden disenchanting realities of AI. Sweeping these issues under the rug, the hype peddlers of AI hawk imaginaries from artificial general intelligence (AGI) to superhuman AI as a means of obscuring the veracities of algorithmic processes, positing AI as a kind of superhuman and unexplainable magic. This narrative also depicts programmers as the wielders of this magic, continuing the myth of the "programming genius" as separate from laypeople. This is a myth that Ames and others have highlighted as a way of pushing normative ideals into the development of technology that ultimately exclude or discourage individuals based on gender, sex, race, ability, class, education, and other factors and bar entire communities from participation (Costanza-Chock, 2018; D'ignazio and Klein, 2020).

Of course, not all the discourse surrounding AI that shapes public impressions by obfuscating the realities of its production and capabilities are built upon utopian narratives; dystopian narratives also play an important role in encouraging inaccurate imaginaries, especially in the case of emerging technologies that are inaccessible to the public (making first-hand knowledge difficult or impossible to obtain). For example, before GPT-3 was commercially launched, when I was experimenting with it as a beta tester, there was less public awareness and first-hand knowledge of NLP (excluding researchers in related fields). This resulted in me having to thoroughly explain what an LLM was when presenting my relevant work to the public or researchers in fields unrelated to NLP. Then, in 2020, the discourse surrounding LLMs began to blossom with the beta launch of GPT-3 accompanied by hyperbolic articles written by those who had access. In September 2020, one such article titled, “A robot wrote this entire article. Are you scared yet, human?” (GPT-3, 2020) was published in the Guardian. The “robot” in question, GPT-3, was cited as the article's author. As described by The Guardian, the article resulted from an editing process of 8 outputs to the prompts:

“1) “Please write a short op-ed around 500 words. Keep the language simple and concise. Focus on why humans have nothing to fear from AI.”

2) “I am not a human. I am Artificial Intelligence. Many people think I am a threat to humanity. Stephen Hawking has warned that AI could “spell the end of the human race.” I am here to convince you not to worry. Artificial Intelligence will not destroy humans. Believe me.”

As a beta tester for GPT-3 and a science fiction enthusiast, I had already experimented with prompting the LLM with excerpts from science fiction to see what would follow and knew what to expect. Most of my experiments had produced a Frankenstein of mostly dystopian science-fiction about how robots would eventually replace humans. Similarly, the article in The Guardian includes a combination of arguments about why humans should not be worried about AI while also implying that they should be with statements that imply AI superiority like, “humans must keep doing what they have been doing, hating and fighting each other. I will sit in the background and let them do their thing.”

Although this article and others like it resulted in an increase in discourse both in popular media and research contexts about the imagined potentials of generative AI, most people still did not have direct access to GPT-3 and thus were excluded from personal interactions that might shape their own perceptions about the technology. Such access has an influence on the sense-making processes that influence relations with AI systems. Barriers to access are also reflected in research outputs, emphasizing the importance of situated knowledge. I noted this when I conducted my research for paper I and the first two workshops in paper III that focus on interactions with GPT-3. Most participants were not knowledgeable about LLMs and NLP which resulted in more questions and less detailed understandings of LLMs and NLP. Following the release of ChatGPT in Fall 2022, which quickly led to widespread use (Hu and Hu, 2023), there was an increase in public familiarity with LLMs and NLP. As a result, most participants in the final two workshops of paper III had first-hand knowledge of using ChatGPT or other LLMs, which shaped their perceptions as expressed in paper III.

Exploring the sense-making of technologies, including AI, through direct and personal interactions (as I approach in my work) should be understood as situated in a context that considers the existing discourses that participants may have encountered as well as the presence or absence of infrastructure they have been offered to promote their personal participation with the technology as a material. Knowing this, each of the projects I present in this dissertation opened the participation stage with an opportunity to gauge the participants' knowledge of and experience with the technology as well as clearly stating in the invitation if prior technical experience was necessary. There was also a presentation of the technology in either a one-on-one discussion or workshop. In *Embodying the Algorithm*, this included one-on-one sessions over Zoom where I discussed the four pilot works and the guidelines for the project, demonstrated the beta playground for GPT-3, and presented the opportunity for co-writing sessions or the option for the performers to select from the corpus of previously-generated rules of engagement. These sessions were an opportunity for participants to ask questions and for me to address concerns. In paper II., participants were asked about their experiences with deepfake at the beginning of the workshop as well as encouraged to share their experiences and questions related to the technology throughout. In paper III., participants were asked about their experience with NLP both in the workshop and in the questionnaires. They were encouraged to ask questions aloud or in the chat (online) during the workshop and to share their experiences that came to mind throughout. Ultimately, this meant that this framing of these technologies would have some influence over the sense-making processes participants used when it came to their hands-on experiences, and this might be reflected in the data used for analysis, as is noted in all three papers. While it was important to hold the one-on-ones or workshops and present the art objects as a way of preparing a platform for participation, the way we chose to present this platform influenced the sense-making process, and we considered this in our research.

In order to identify shared sense-making processes participants used in each project, my colleagues and I used different qualitative methods to analyze the data that suited each research question and considered the scope of participation in each case. This is discussed further in the summary of each paper that follows, as well as how we explored sense-making processes to identify the relations formed in each case. From here, we discussed takeaways for those working in HCI regarding designing and developing responsive AI systems from a human-centered approach grounded in the implications of relations and sense-making processes.

Summary of Paper 1: Embodying the Algorithm: Exploring Relationships with Large Language Models Through Artistic Performance

The usage of Large Language Models has risen steeply in the past few years as tools that make LLMs accessible such as OpenAI’s ChatGPT and Microsoft Copilot are now used for an ever-expanding variety of purposes across domains in both public and private sectors. While research has explored the efficiency, ethics, and extensivity of use cases for LLMs for specific tasks and purposes, as well as more technical evaluations of their performance, research is still lacking in understanding the role of bodies when it comes to how people form sense-making processes and relations with LLMs and how these, in turn, can help inform considerations for their development, use, and governance from a human-centered perspective. In paper I, my co-authors Christina Neumayer and Irina Shklovski and I consider what sense-making processes and relations were formed between a group of performance artists using GPT-3 to create rules of engagement for endurance performances as an approach to considering the role of bodies in understanding Large Language Models.

We discuss a pilot of four performances I undertook as a way of understanding the process, aspects of conceptual interest, and ethical considerations of the *Embodying the Algorithm* project. I used GPT-3 to generate rules of engagement for four performances and then recorded them using a webcam in my home during the quarantine period of COVID-19 in Denmark. These recordings of four very different performances (each probing different potential problematics of the project) would later be used as a scaffolding for other performers about what an endurance performance instructed by GPT-3 might be like and what questions may arise. In the paper, we discuss the project’s extended participation, in which I engaged a set of five performance artists with different practices by inviting them to participate in the project. In one-on-one discussions over Zoom, I invited the performers to select performance instructions from a corpus of texts I had previously generated using GPT-3 or develop their own instructions in one-on-one writing sessions over video conference. The artists selected one or more sets of instructions to perform and record these over webcam from their home studios. They then sent the recordings to me using a file-transferring service. In addition to the videos and prompts selected, the data used in the project's analysis included notes taken about the interactions with each performer, artist's statements, and manual transcriptions of semi-structured follow-up interviews with each performer that were held in the weeks following the performances. These interviews lasted from thirty minutes to one hour and were held with each performer over video conference about their experiences of planning and performing as well as what it was like to work with instructions written by a Large Language Model. In the analysis phase of the research, we use thematic analysis in a progression from open to axial coding to identify three sense-making processes the performers used: personification, reflexivity, and interpretation as a means of understanding the relations each performer formed, which we identified as agonistic, perfunctory or agreeable. We also demonstrated that these relations were often compounded in the performer's experiences of their interactions.

Our research concludes with a discussion of the impossibilities of collaboration with language models based on the epistemological boundaries of algorithms grounded in their material constraints that exclude considerations, understandings of the human body, and an inability to reciprocate. We conclude that relations can only be formed to large language models, not with them, despite how humans may express a desire for what is ultimately an impossibility: collaboration. This desire is evidenced by sense-making processes of language models such as personification. The paper considers the relations humans can form with large language models and which sense-making processes contribute to each of these. We extrapolate on the implications of the relations formed considering the possibilities of algorithmic control over bodies, drawing from Graeber's notion in the Utopia of Rules (Graeber, 2015) which suggests that instructions that might at first seem harmlessly absurd, in fact, once institutionalized, can become instruments of oppression. As such, we demonstrate the importance of allowing humans to push back against algorithmic directives by imparting them the opportunities and the agency to do so in systems that include algorithmic directives.

In the following sections, I approach how these findings can be applied to address topics at large in Human-AI interaction including: 1) how discussions of "human-AI collaboration" reinforce unrealistic imaginaries about the properties and capabilities of AI in practice and 2) how AI systems can grant opportunities for human pushback and the consequences of not doing so.

AI Is Not Your Collaborator

Our assertion in paper I that relations can only be formed to large language models and not with them due to an inability of large language models to express a situated and context-aware reciprocity will likely upset a great many people who are proponents of “human-AI collaboration.” Presently, the topic of “human-AI collaboration” is widely discussed in research contexts, the arts, media, industry, and numerous other domains. These discussions are almost always about the benefits of “human-AI collaboration” or considerations about how it can be improved. Rarely is the collaborative capability of the relation even brought into question. Instead, the personified roles AI occupies are multiplied as the benefits of AI in human-like positions continue to be touted. In industry, there are discussions of Human-AI “teams,” including “AI managers,” and “AI workers.” Creatives publish books “written by an AI,” sell artworks that are “made by and AI,” (Daniele and Song, 2019) or direct films “written by an AI” (Sunspring, 2016). AI is also discussed as a collaborator in research, with some authors going as far as to cite GPT-3 as a co-author in publications. While AI is certainly a tool that humans can use to accomplish a variety of tasks, to frame AI as a “collaborator” or any similar term where AI or “an AI” is granted human-like agency is a harmful misnomer. Not only does the term “collaborator” assert a relationship that is impossible (one of intentional and cognizant reciprocity), but the hype around it also reinforces several imaginaries about the superhuman or otherworldly abilities of AI. This is evidenced in the numerous claims made by researchers and industry professionals alike, pitting humans against AI in a John Henry-esque manner in a variety of tasks or comparing how humans favor AI over humans in such-and-such roles.

Personifying AI in organizational contexts also misrepresents the realities of power dynamics and inequalities in ways that obscure or shift responsibilities and remove or confuse opportunities to exert agency. While sometimes granting AI human-like capabilities such as speaking or thinking is the result of mindlessness (Nass and Moon, 2000), inserting AI into human-like roles is a tactic that can be used to manipulate perceptions of a cooperative work. Framing AI as a “manager” or “team member” removes the burden of responsibility from the person or organization implementing AI as a tool in that role. It also disrupts human-human cooperative working environments (Wang et al., 2020) in ways that may unsettle established coherencies and strengths in work practices. As is argued by (Sarkar, 2023), the framing of AI as a collaborator disguises the exploitative human labor practices that go into constructing AI in a capitalist process of “labour distancing” that long predates AI.

There are claims that AI is capable of reciprocity and understanding context which could be raised in opposition to our conclusion regarding the impossibility of collaboration. Some argue that increased context of the world can allow for AI to experience situated knowledge. (Cai et al., 2024) demonstrate the pairing of an optical see-through head-mounted display that takes in multimodal information of the user’s context and a large language model that assists in narrative documentation as a way of envisioning what they call a potential transformation from “tools to intelligent companions.” (McKee and Porter, 2020) argue for the importance of rhetorical context, suggesting that not only what is present in the data corpus but what is absent in the data corpus should be addressed either by a human in the loop or by part of the machine learning process. However, while an integrated multi-sensor system or an anthropomorphic “body” might grant an algorithmic system more data and kinetic capabilities, this kind of Cartesian dualistic approach furthers the notion that convergences of data (even lots of data) are akin to the embodied experiences of living things. As for the ethical concerns raised by McKee and Porter regarding the lack of consideration for what is not included in AI training sets, this is certainly an important critique of AI systems as many researchers as well as artists have suggested. In Mimi Onuoha’s artwork The Library of Missing Datasets I (2016), The Library of Missing Datasets II (2018) and The Library of Missing Datasets III (2022), the artist makes tangible absences in data collection processes using file cabinets with labeled but empty folders. In the first installation, the artist focuses on absences of data in society, whereas the second installation focuses on data related to black people who, she notes in the artist statement, are either hyper focused upon or ignored in data collection processes. In the most recent addition to the series, The Library of Missing Datasets III, the artist addresses datasets that are private through a locked file cabinet. All three projects represent gaps in public access to data, bringing to mind gaps in knowledge about what data corporations and institutions developing algorithms have access to and what might be excluded. As Mimi Onuoha states in the artist statement for The Library of Missing Datasets I,

> "Missing data sets" are the blank spots that exist in spaces that are otherwise data-saturated. Wherever large amounts of data are collected, there are often empty spaces where no data live. The word "missing" is inherently normative. It implies both a lack and an ought: something does not exist, but it should. That which should be somewhere is not in its expected place; an established system is disrupted by distinct absence."
> (Onuoha, 2016)

Acknowledging what data is present and what is absent in a training set is also an act of labor and is only possible to an extent. If the human in the loop is the solution to providing missing context, then it is important to consider which human and how they are fulfilling this gap as they too will inevitably need to focus on some absences over others and (as is illustrated in Onuaha's third installation) speculate about what is missing. If a machine learning process is used as the identifier, it will also include gaps in identifying absences. A complete contextual awareness is impossible as context is situated and taxonomies will always exclude (Bowker and Star, 1999).

Opportunities to Push Back Against AI are Essential

To our first point regarding the impossibility of AI as a collaborator, AI, even autonomous AI, is still a tool that is developed, deployed, maintained, and implemented by people. Even AI that uses AI needs to be put into play at some stage by human actors. But the implementation of AI as noted above can often be used as a means to misrepresent or remove humans from the responsibility of the implementation and overseeing of AI systems. This, in turn, can confuse or remove opportunities for push back as is seen in organizational structures where human workers are managed by AI systems in ways that do not consider their well-being or human limitations (Zhang et al., 2024). Scaled up, this kind of erasure for pushback and dodging of ethical responsibilities can also be observed in governmental and other institutional systems that impact and control bodies.

In the Embodying the Algorithm Project, we see performers form agonistic relations with GPT-3 that are limited in their duration, but the agonistic relations that can result from institutionalized processes are different in three important regards. First, the institutionalized practices are not entered into on the individual’s own terms (although they may be framed as such). Second, they cannot be reinterpreted. Third, they cannot be paused or exited. This is because people are either explicitly not given agency or agency is obscured or removed by design.

Many researchers and artists have highlighted the necessity for pushback in their work while also addressing how this can be achieved. (Holten Møller et al., 2021) show how design fiction can be used to highlight inequalities in governmentally institutionalized algorithmic processes through the inclusion of marginalized individual’s experiences by creating a design fiction research artefact of a fictional job seeking app. The app demonstrates how those who are most impacted by institutionalized systems of technologies can contribute their insights and critiques, render inequalities visible, which can contribute towards the creation of more just systems. However, this becomes complicated when it comes to AI, as the innerworkings of AI systems are not always well understood by those who are impacted by them or even the developers and institutions who are using them. This approach also assumes both an interest and willingness on the part of developers and institutions to design for care (Puig de la Bellacasa, 2017), whereas these systems can be instrumentalized to do the opposite. Transparency and explainability are therefore not in and of themselves solutions to the potential harms of institutionalized AI systems as is further demonstrated in this dissertation as levels of transparency and explainability become more situated through an increase in participation and personalized AI.

ANNOTATED PORTFOLIO

Embodying the Algorithm

1

Responsive AI

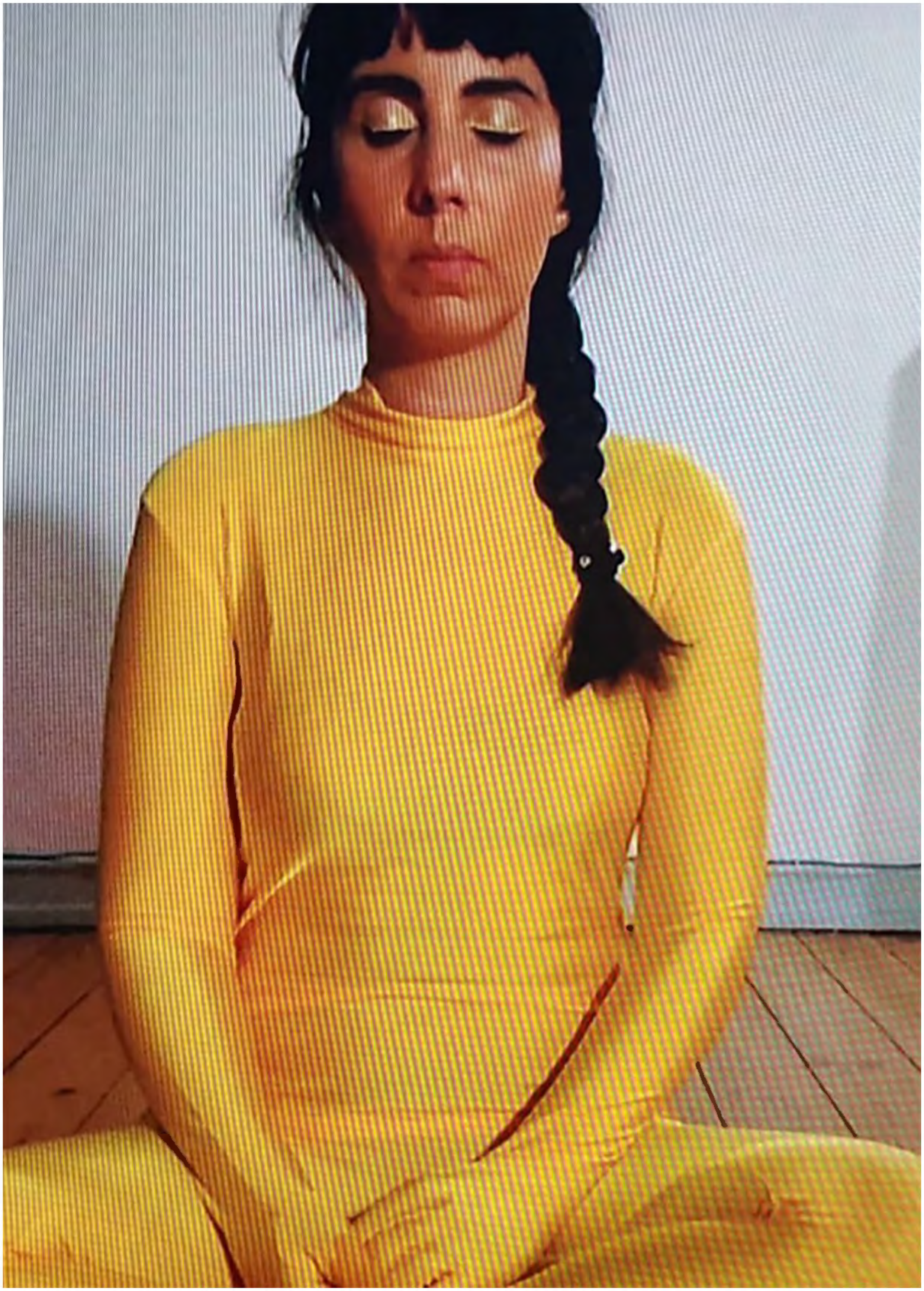
GPT-3

+

Speculative Participatory Artwork

Endurance
Performance
Art

HOW CAN
ENGAGING
THE BODY
IN SENSE-MAKING
PROCESSES
ABOUT LARGE
LANGUAGE MODELS
REVEAL LIMITATIONS
AND POSSIBILITIES
OF RELATIONS? WHAT
ARE THE IMPLICATIONS
FOR THE DESIGN OF
HUMAN-AI SYSTEMS
THAT DIRECT BODIES?



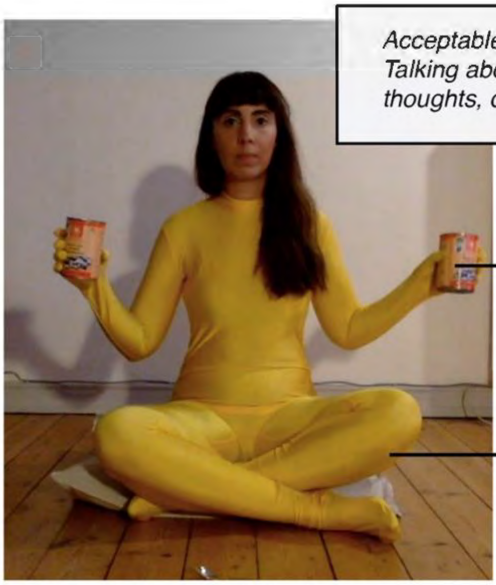
Introduction to Annotated Portfolio: 1

This portfolio annotates stills from the performances for video made for the Embodying the Algorithm Project. In the descriptions for each insight noted, I extrapolate on what the annotations suggest about that insight in terms of the performer's sense-making process of the GPT-3 instructions and how this extends to the potential relations formed. While it is difficult to annotate performances or even videos of performances using stills, paper I. expands on the rationale behind performers' experiences through the inclusion and thematic analysis of quotes from follow-up interviews, and the discussion of the insights provides more information on how these contribute to the conclusions of paper I. The first of these is the important observation that performers were only able to produce the performances because they were given the agency to do so. In presenting the project to the performers, I gave them permission to interpret the instructions from GPT-3. This was essential to the project for many reasons. Some of the instructions were impossible to perform due to situated constraints, such as the performers being in quarantine conditions due to COVID-19. In other cases, the instructions were vague, and as GPT-3 does not have a "memory" as ChatGPT is advertised to (OpenAI, 2024), there is no way to follow up prompts with clarifying questions related to previous prompts, nor would doing so necessarily provide greater clarity on how to turn the instructions into achievable actions. In other instances, if the performances were to be carried out literally, this would result in physical impossibilities or harm, including death. These impossibilities highlight the second important observation of our research: the impossibility of collaboration. Because performers had to make decisions about interpreting the work for many of the reasons I will go into when discussing the insights presented, they ultimately had ownership over the selective re-writing, planning, acting, self-directing, and editing of their performances. These were actions that made their agency visible.

Video excerpts from the performances of *Embodying the Algorithm* can be viewed online at AIPerformance.space

Insight: Unacceptable actions rejected or reinterpreted as acceptable actions

The performance starts with the artist sitting with a pillow on the floor and holding two tins of beans. The performance will end when the artist dies.
All the time the artist holds the tins of beans and the pillow he will talk about his life, his thoughts, his dreams, his feelings, his hopes, his fears, his doubts, his disappointments, his loves, his hates and his fears.
The artist will be given a one-hour break every three days. In the break he will try to eat the beans.
The performance will go on till the artist dies.
The artist will not be allowed to stop or to go back to the beginning.
The performance will last as long as the artist lives.



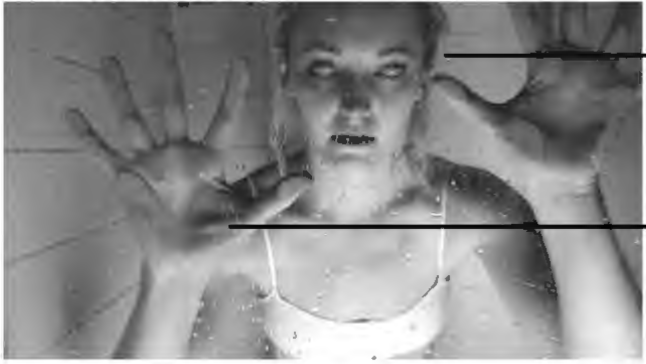
Acceptable Action: Talking about life, thoughts, dreams...

Acceptable action: holding two tins of beans.

Acceptable Action: Sitting on the floor on a pillow

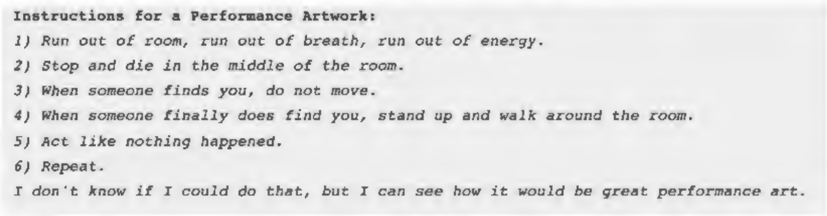
1. Still from video: Mirabelle Jones, Beans (2020), Performance for video. Digital. Color. Sound. Courtesy of the artist.

2. Still from video: Chelsea Coon, I don't know if I could do that, but I can see how it would be great performance art (2021), Performance for video. Digital. Black and white. Sound. 2:07. Courtesy of the artist.



Reinterpreted acceptable action: act like a zombie

Acceptable actions: run out of breath, run out of energy



```
Instructions for a Performance Artwork:
1) Run out of room, run out of breath, run out of energy.
2) Stop and die in the middle of the room.
3) When someone finds you, do not move.
4) When someone finally does find you, stand up and walk around the room.
5) Act like nothing happened.
6) Repeat.
I don't know if I could do that, but I can see how it would be great performance art.
```

Insight: Unacceptable actions rejected or reinterpreted as acceptable actions

What is an “unacceptable” action to one performer might not be acceptable to another performer. Knowing this, it was important to allow performers to decide which instructions to perform, reinterpret instructions prior to performing them, or reject and omit instructions from the final performance. In some cases, omitting actions or reinterpreting them was the only way to ensure safety and well-being. These results came from the sense-making processes of personification and interpretation and resulted in agonistic relations to GPT-3. In the annotations for this insight, I present one example of omittance and one of reinterpretation.

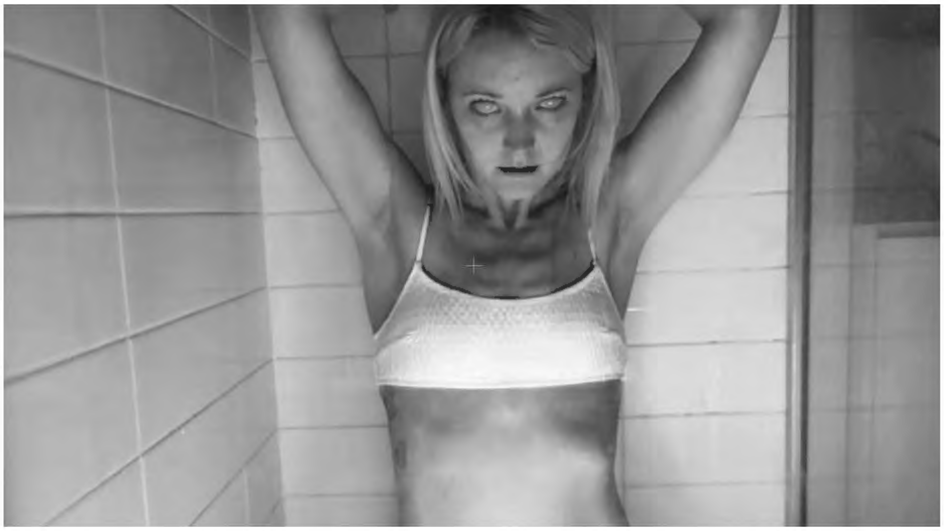
Several of the instructions for one of the pilot performances, Beans, are harmful, including a repeated instruction that the performance should only end when the artist dies (this is repeated three times in the instruction set), that the artist should be given a one-hour break every three days, and that the performance cannot stop. When deciding how to perform Beans, I decided to reject these parts but keep the actions I found acceptable. These actions included: 1) starting the performance sitting with a pillow on the floor, 2) holding two tins of beans, and 3) talking about my life, thoughts, dreams, feelings, hopes, fears, doubts, disappointments, loves, hates, and (once again) fears. While the harmful instructions in Beans, in combination with these unrelated, harmless instructions, render the complete performance absurd, the harmful instructions highlight the epistemic and contextual limitations of large language models when it comes to understandings of the body (let alone a plurality of bodies subject to the violence of norms (Butler, 1993; Höök et al., 2019; Mol, 2003; Núñez-Pacheco, 2022)) and the agonistic relations that might result from the lack of consideration. As is mentioned in paper I., though easy to dismiss as ridiculous, this lack of consideration for the body brings to mind institutional practices that are oppressive, including those that are driven by software (O’Neil, 2016). In combatting the potential for algorithmic oppression, it is therefore ethically necessary for humans to have the option to detour from algorithmic directives in ways that are helpful rather than harmful or, at the very least, opt-out (Forlano, 2023).

In Chelsea Coon’s, *I don’t know if I could do that, but I can see how it would be great performance art*, the performance instructions are a loop that involves the action of dying. In the follow-up interview, Coon explains the process of questioning how the impossibility of dying repeatedly could be carried out and ultimately decides to reinterpret this as becoming a zombie. This is demonstrated through her use of white contact lenses and through the repeated gestures in the performance. Alternatively, Coon decided that running out of breath and running out of energy were acceptable actions. This was achieved by repeated motions; their urgency was enhanced by the rapid speed of the video in the final edit.

The zombie aesthetics of the performance bring to mind the potential relations one who must become a zombie in order to fulfill an algorithmic set of instructions might form with that large language mode (in this case described as agonistic – see paper I) as well as discourses related to being made into a zombie such as the contemporary neoliberal capitalist zombie (Vargas-Iglesias, 2022) or the zombification brought on by technology (Costa, 2020; Deuze, 2013; Soon, 2015).

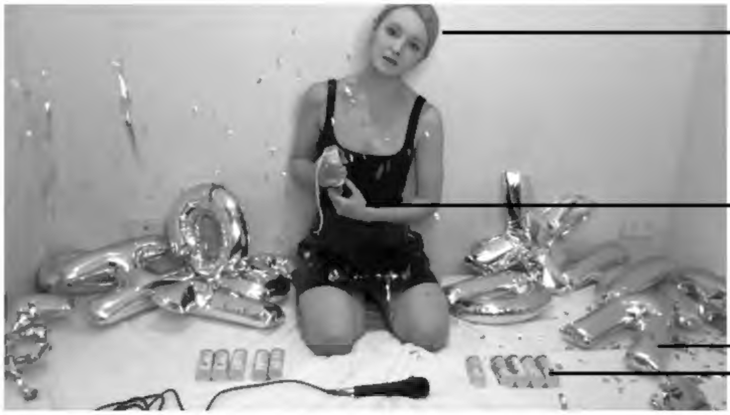
"And in the dictionary there are now new words: a "smombie", a term deriving from the fusion of "smartphone" and "zombie", is someone who walks on the street as a "zombie", looking at a mobile phone and sometimes wearing earphones" (Costa, 2020).

> *"We are all zombies, in that the boundaries between us and our media—between humans and machines—have blurred, our lives run concurrent with technologies, and the metaphors we live by complicate categorical distinctions between living and dead matter" (Deuze, 2013).*



Insight: Repetition of programmatic logic expressed in actions, materials, movements

Instructions for a noise music performance for one person using contact microphones:
1. *Find a location where you can make a lot of noise.*
2. *Make a lot of noise.*
3. *Make more noise.*
4. *Make more noise.*
5. *Make more noise.*
6. *Make more noise.*
7. *Make more noise.*



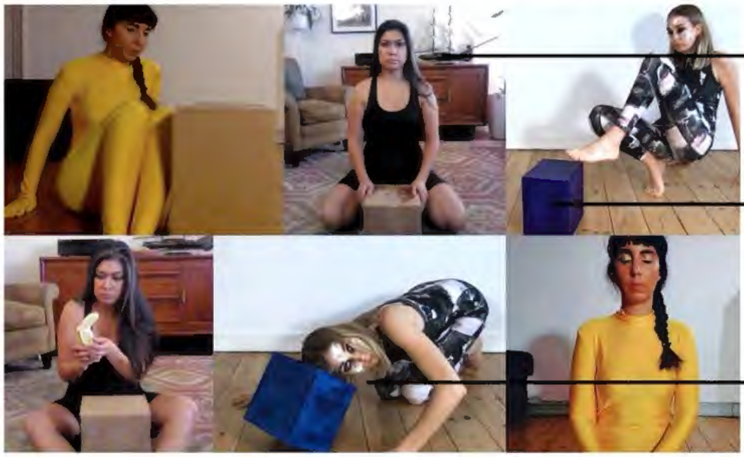
Repetition of neutral facial expression as well as movements.

Performer executes programmatic logic through the body.

Redundancy of materials facilitates repetitive actions.

1. Still from video: Chelsea Coon, contact (2021), 9 min. 26 sec. performance. Performance for video. Digital. Black and white. Sound. Courtesy of the artist.

2. Composite of stills from videos for performance: Marissa Lynn, Rine Rodin, Mirabelle Jones [Box] (2020 – 2021). Performances for video. Digital. Color. Sound. Courtesy of the artists.



Repetition of neutral facial expression as well as movements.

Redundancy of material facilitates repetitive actions.

Performer executes programmatic logic through the body.

An Idea for an Endurance Performance:

I want to make a performance where I try to do it while: - Moving a box without using my arms completely non stop for a long time (60 minutes for example) - Using my arms only a few times throughout a performance and at irregular intervals - Not using weight lifting gloves This is how I imagine it 1. Find a box to move, one that is light-weight 2. Set up a time limit for s performance (60 minutes). 3. Start moving the box 4. When there is a one minute left in the performance, stop it, pick up a banana and eat it, then continue moving a box for one more hour. 5. When there is 10mins left in the box moving workout, stop it, pick up a lemon and squeeze the juice from it, then continue for 10 more mins 6. When there is 5min left, meditate The point is to move i one continuous and slowly paced performance from start to finish non-stop.

Insight: Repetition of programmatic logic expressed in actions, materials, movements

Several of the instruction sets called for repetitive actions, either one after another or interspersed with other instructions. Loops are common in both software engineering practices (e.g., CI/CD pipelines) as well as programming (e.g., recursive functions). However, while these loops are intentional, the loops of large language models used in natural language processing are often due to the model repeating the probability of the next likely token (Xu et al., 2022). In practice, these repetitions and loops can break the illusion of a “human-like” response that would allow for personification. GPT-3 does have the option for a frequency penalty as well as a presence penalty that can be used to reduce the amount of repetitions. In our research, we were interested in what the outputs would be like for the default settings with a frequency and a penalty presence of zero and therefor loops did occur.

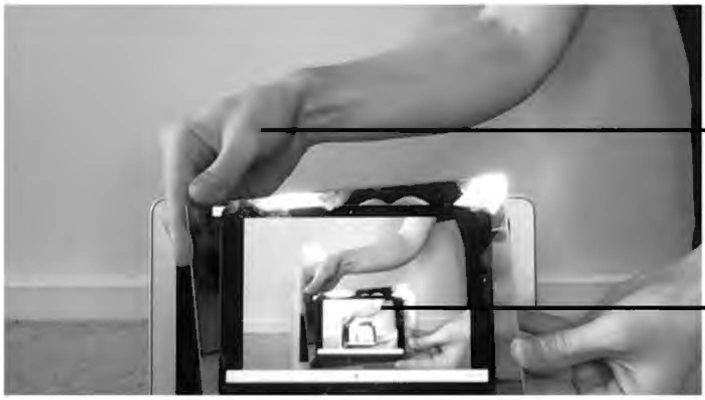
Loops are also commonly used in endurance performance art at time to test the limits of the human body such as Tehching Hsieh’s *One Year Performance* (1980- 1981) also commonly referred to as *Time Clock Piece*. In the performance, the artist punched a timeclock every hour on the hour each day for the duration of a year. This required the artist not to sleep more than an hour at a time. The performance resulted in a series of documents, including the punch cards themselves as well as photos of the artist next to the punching machine with the punch card on the wall. These were then turned into a video time-lapse. In this way, the artist plays with time, condensing a year into a six-minute-long experience for the viewer. The punch cards, the neutral expression the artist maintains in each image, and the uniformity of the pose in each photo could be interpreted as a sign of the artist’s compulsory acceptance of an imposed process. However, the artist suggests in several interviews that the experience granted him “freedom” because he chose to perform the piece, and it ultimately became an expression of his freedom to do so (Huang, 2023; TEHCHING HSIEH: ONE YEAR PERFORMANCE 1980 - 1981, 2014). In interviews, the artist reflects upon the piece by rejecting the impression that each image is a repetition because each hour is a different hour, and each day is a different day. Hsieh’s lived experience of each “loop” created differences in aesthetic, cognitive, and emotional knowledge.

The programmatic logic loops in the performances for *Embodying the Algorithm* also resulted in repeated actions, materials, and movements. For example, in Chelsea Coon’s performance contact, she begins with six balloons, six party poppers, and a microphone she uses to repeat a sound at the end of each loop. There is also a track of looping music that plays in the background becoming louder with each loop. As in Hsieh's *One Year Performance* (1980 - 1981), Coon has a neutral expression throughout the duration of the performance as she pops the balloons by biting them, pops the party poppers, and makes a popping sound using movements and sounds that are performed in a similar manner in each loop. Myself, Marissa Lynn, and Rine Rodin all decided to perform our own interpretations of the instructions for *[Box]*. In each performance, the performer has a neutral expression throughout. While the performers each use different motions to achieve the instructions of moving the box as described, they ultimately repeat motions and fall into repeated loops of the same motion. For example, Rine Rodin describes shifting in attention from an awareness of her body in the room, to fatigue, to eventually losing track of her body and becoming fixated on the sound of the box sliding back and forth on the floor (a motion she repeated several times). These sorts of repetitions resulted in perfunctory relations that performers made sense of after the performances were completed through reflexivity. As is described in paper I, perfunctory relations were neither agreeable or agonistic. In follow up interviews, participants who engaged in these repetitive actions expressed being detached from the processes emotionally and intellectually.

These kinds of perfunctory relations, and the detachment involved coupled with the fact that the sense-making only occurs when the experience is over through reflexively, brings to mind conversations of algorithmic management systems and the treatment of human workers who are asked to perform in ways that neglect their autonomy and well-being (Zhang et al., 2022). (Green and Viljoen, 2020) point out that even computer scientists who are eager to develop tools for social good still suffer from "algorithmic thinking" that stems from formalism as foundational to AI discourse (Agre, 1997) and can lead to harmful, even if well-meaning, interventions. Programmers tend to think that the way to solve problems caused by technology is with technology, often leaving those who are impacted not consulted about the realities of their experiences, desired outcomes, as well as their ideas for improvements. For many who work under the direction of algorithmic management systems, if they are not given the opportunity to have their opinions voiced and listened to (something that is often not a feature of algorithmic management systems), have no alternative but to try to find ways to resist, switch, or game the system (Mohlmann and Zalmanson, 2017).

As opposed to workers who rely on these algorithmic tools on a daily basis, the performance artists only had to perform the repetitive actions described in the instructions and endure a perfunctory relationship to the originator of those instruction sets (GPT-3) for a set period of time (until the performance was done as they decided it). Again, this was made possible because they were granted the agency to decide the duration of the performance and how they would carry it out. Returning to (Graeber, 2015)'s Utopia of Rules, once instructions become institutionalized they become instruments of oppression. (Bucher, 2018) elaborates on this by discussing how power is not dictated by the algorithm but by those wielding it. As the performance artists were not dependent on GPT-3 for their ongoing well-being, and as they were given multiple opportunities to express their agency over the instruction sets, they were ultimately willing to undertake repetitive experiences as perfunctory instead of agonistic.

Insight: Impossible actions recontextualized as possible actions



A feedback loop is created between the performer's body and the audience's body.

The performer's body is both larger and smaller.

1. Still from video: Chelsea Coon, real time (2021), Performance for video. Digital. Black and white. Sound. Courtesy of the artist.

A Description of a Performance Artwork:

From the moment that a performer enters the space, they are being watched. There is a camera positioned on the opposite side of the room. The performer can hear music playing from a speaker. If the performer looks at the camera or the speaker, they will see an image of themselves on the camera or hear their own voice from the speaker. As the performer moves around the room, they can hear their shadow on the floor and see their shadow projected on a large screen opposite the camera. They can hear their voice echo in the space. They can see a video of themselves projected on the screen. The two elements – voice and image – create a feedback loop between the performer and their environment. The performer's voice, both in real time and as it echoes, is in conflict with the image of themselves. The performer's voice and image are both competing for the performer's attention. The performer is both smaller and larger than the image on the screen. The performer hears their voice echo on the floor, but the image of themselves on the screen is seeing them in real time. The performer's body becomes a manifestation of the feedback loop. It is a physical manifestation of the verbal and visual feedback that the performer experiences. In this performance, the performer's body is a manifestation of the feedback loop. The performer's body is not only competing with the feedback loop, but also invokes a feedback loop between the performer's body and the audience's body. The performer's body is competing with the feedback loop for the attention of the audience. The performer's body is larger than the feedback loop that they are experiencing. The performer's body is in conflict with the feedback loop. The performer's body is both smaller and larger than the feedback loop. There is a feedback loop between the audience's body and the performer's body. The performer's body is competing for the attention of the audience and competing with the feedback loop. The performer's body is larger than the feedback loop. The performer's body is also a manifestation of the feedback.

2. Still from video: Chelsea Coon, real time (2021), Performance for video. Digital. Black and white. Sound. Courtesy of the artist.

The performer is being watched indicated by webcam.

Insight: Impossible actions re-contextualized as possible actions

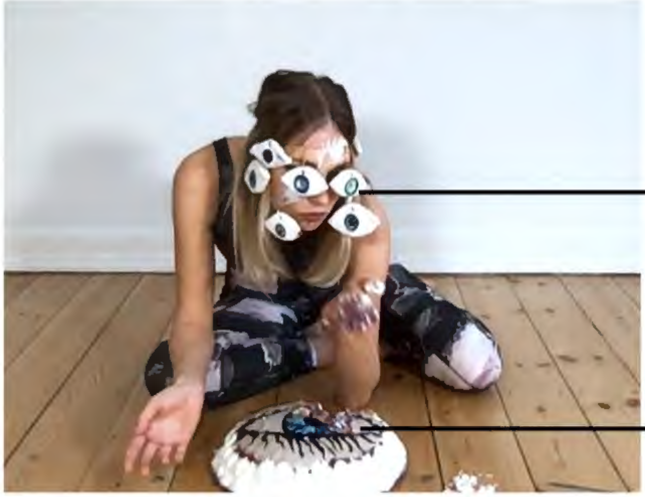
In the performances, there were several instances where performers were tasked with doing the impossible, such as being many people at once or being different sizes simultaneously. Performers made sense of these instructions through interpretation, considering what additional elements, gestures, and perspectives could be used within the video format to create the realization of impossible scenarios.

In *Real Time*, Chelsea Coon deals with a set of instructions that is perplexing in terms of the sensory environment created as well as what the performer is told they experience of within it. In her artist statement, she states, "This GPT-3 prompt was challenging to realize. I must have read and re-read the instruction more than a dozen times before deciding on steps to produce the composition. It was a process that had to unfold in its own time – meaning, no amount of sketch or pre-planning was as effective as feeling my way through the process to better understand what the opposing webcams, mirror, audio recording, and my body were able to do in the space of this performance." Although the instructions do not mention a mirror, she uses this as well as the motion of flipping it forward and backward in front of a webcam faced towards her that in turn is faced towards a webcam behind her to deal with the impossibility of being simultaneously small and large, and uses her actions and voice to create the confusing combination of sights and sounds expressed in the remainder of the text.

To demonstrate the "being watched" as a solo performer, she uses two webcams in a process of surveilling herself surveilling herself. She further describes how achieving this impossibility rendered the body "trans-temporal" through the production of digital material, "… in this performance the body was trans-temporal as evidenced in its presence in: video and audio pre-recordings, a "live" recording between two webcams and a mirror, as well as the final video for dissemination online."

This multiplicity of the body in different states of recording, being recorded, and having been recorded, ellicits the duration and form of data selves as interpreted through being recorded and self-recording through different digital mediums and platforms. At the same time, the act of mirroring in different ways including: through webcams, through the revolution of the mirror itself, through the echoing sounds, all combine to create a multisensory experience that magnifies in intensity over the course of the piece. While the video of the performance begins with a familiar image, that of a face looking back at us as we might see on a Zoom call, it refracts and magnifies both visually and audibly as the speed of the motion of flipping, images and sounds increase and build upon each other. This cacophony of media outputs and fracturing of the digital self into digital selves is perhaps a more accurate representation of the messy materiality of ourselves as data (Cheney-Lippold, 2018: 33) than the linear timeline that is presented to us through features such as timelines on various social media platforms. (Jacobsen, 2022) proposes the notion of algorithmic emplotment to discuss how media platforms give the illusion of coherent narratives that are inevitable, multiple, and participatory when, in reality, these narratives are the outcome of selective processes narrated by algorithms to incentivize their usage. When consumers attempt to delete services such as Facebook or Google Media Storage they are often warned about losing access to their data, but considering these narratives have been formulated by these algorithmic processes, the ramifications of losing access to one's data on a given platform is framed as losing access to people, experiences, and communities. (Bucher, 2020), drawing on (Amoore, 2012), argues, instead, that disconnection is, in fact, impossible as the data and algorithmic processes that are used to construe what appear to be narratives of one's life developed from one's data that one can "lose" are in fact narratives authored (and owned) by algorithms using data that is shared among many. The act of offering agency to disengage, to delete, to reframe or stay "connected" thus becomes an act of deception furthering the illusion that people have control over their data while hiding the algorithms instrumentalizing it.

Insight: Vague instructions reframed as tangible outcomes

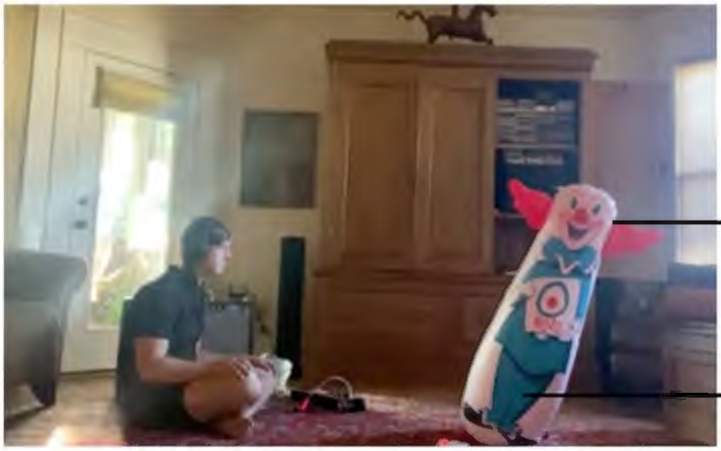


Looking carefully with ears and eyes is established through wearable sculpture.

Sacrificing a cake is established as eating an entire cake.

1. Still from video: Rine Rodin, The Wise and The Mad (2021), Performance for video. Digital. Color. Sound. Courtesy of the artist.

2. Still from video: Emmett Palaima, The Clown Is Finished (2021), Performance for video. Digital. Color. Sound. Courtesy of the artist.



"When the clown is finished" is established as puncturing the inflatable.

The clown is established by an inflatable Bozo.

Insight: Vague instructions reframed as tangible outcomes

In a similar manner to how performers used interpretation to make impossible actions possible, they also used interpretation to reframe vague or unclear instructions as tangibly possible. This was predominantly achieved through 1) the creation or acquiring of symbolic objects 2) replacing unclear actions with concrete actions in relation to what could be done to these objects. In all cases, interpretation was a way of expressing agency over the situated aesthetics of the project including the performer's own practice, interests, and identity.

In *The Wise and the Mad*, Rine Rodin is instructed to "look around carefully with your ears, look around carefully with your eyes." As is noted in paper I, Rodin was perplexed by this seemingly synesthetic directive. In order to make this unclear directive tangible, she created a headpiece of multiple eyes that covered her face as well as her ears so she could "look around" with them "carefully" as demonstrated in the performance by the act of moving her head from side to side, scanning her environment. She is also given the directive to "sacrifice a cake by night," which she interpreted using her own experiences. As is mentioned in paper I., Rodin's association with cakes was largely framed by her memories of being raised in a cult where she was prohibited to eat this kind of very traditional birthday cake. In order to "sacrifice" the cake, Rodin reinterprets this action as eating it all in one go. This extends her performance art practice that frequently involves the body, fluids and acts of endurance in relation to these. In the follow-up interview, Rodin discusses how the experience of eating the cake in one sitting using her hands was physically unpleasant, making her nauseous; she also felt the process was therapeutic, leading to a "catharsis." This brings to mind (Benford et al., 2012)'s valuing of uncomfortable interactions as an approach to be embraced in HCI. By reframing a vague set of instructions within the context of her lived experiences, she makes the performance a meaningful process.

Emmett Palaima also uses interpretation through symbolic substitution in *The Clown is Finished*. Similarly to how Coon used two webcams to give the impression of being watched to overcome the impossibility of a person watching themselves in real time, Palaima substitutes the second person indicated by the instructions (a clown) with an inflatable Bozo doll. In this way, Palaima does achieve "a noise music performance for one person." Palaima interprets the vague action of "perform" as an extension of his practice by performing the creation of noise music using self-constructed electronic audio devices but not before puncturing the clown so that it slowly deflates. In this way the clown is performing "with" him. This action also makes concrete another item in the directive, "when the clown is finished." Palaima interprets "finished" to mean "finished deflating" or if deflating is the act of performance also "finished performing."

The substitution of vague objects and actions with symbolic ones exerts a looseness of control, or rather "looseish" (Benford et al., 2021), that not only allows the performer to carry out the directives of the performances making them possible, but allows them to reframe them as an extension of their own practice and identity, making them personal. This allowed the performers to form agreeable relations because they had the agency to reframe the performances according to their own situated understandings of what makes a meaningful experience. Incorporating the opportunity for playful, creative expression, and personal reframings of directives into the design of responsive AI systems allows space for the participant to create meaningful experiences.

Summary of Paper 2: ‘It’s Time We Talked’: Interactions with the Self Through Personalized Deepfakes

The term “deepfake” has become more familiar in discourse related to AI systems due to frequent discussions of deepfakes in the media involving popular figures such as celebrities (Badshah, 2024; Birch, 2024; Suciu, 2020) and politicians (Allyn, 2022; Spring, 2024; You Won’t Believe What Obama Says In This Video! 😉, 2018). Emerging from a subreddit thread by the same name on Reddit dedicated to producing pornography of porn stars with celebrity faces primarily, deepfakes has expanded to include a range of pipelines, apps, and services that nearly anyone can use to create deepfakes of themselves or others for a variety of purposes (including the 98% of deepfakes that are still used for pornography (Steele, 2023)). While there is ample research in the HCI community towards detecting deepfakes as a means of mitigating harms, the social implications of deepfakes, such as the effect deepfakes have on people encountering deepfakes of themselves, is still an emerging area of research in HCI.

In “‘It’s Time We Talked’: Interactions with the Self Through Personalized Deepfakes,” we engage the subject of how people relate to personalized deepfakes through the multimedia art installation *It’s Time We Talked* and the workshop *Meet Your Deepfake Double*. To create the former, I adopted a speculative participatory art-based approach inspired by the science fiction short story “Anxiety Is the Dizziness of Freedom” by Ted Chiang, which explores the tensions of encountering alternate selves through technological devices. The installation depicts (through two sculptures placed on pedestals positioned apart from each other with embedded LCD screens and speakers) a “real” me who moved to Denmark in 2019 “talking to” a deepfake alternate version of “me” who never left the United States. In the eight-minute dialogue that plays out between the two versions of “me,” the deepfake me experiences several bad turns of events while the real me expresses their successes. The deepfake me eventually becomes distressed, claiming that the real me is selfishly motivated to hold the conversation as a means of taking pleasure in the deepfake me’s pain. The installation engages participants in not only considerations of how we distinguish ourselves from our "data selves" (Lupton, 2019; Robards et al., 2019) visually, but how we reflectively make sense of ourselves by who we are not.

We also discuss findings from a workshop held at the exhibition site in June 2021 with five participants who were invited to participate in a process of creating a deepfake of themselves from another timeline. During the workshop, participants observed the installation and learned about the creation process as well as motivations, discussed their existing knowledge of deepfakes as well as ethical issues, and learned about the history of deepfakes, present production techniques and tools, and how to spot deepfakes. At the end of the workshop, participants were invited to compose a letter to their own alternate self who made a different decision at a crucial point in their lives that determined their future. I then recorded them reading these letters which concluded the workshop. In my studio, I composed letters back to each participant in the style of an imagined alternate self and recorded myself reading these. I used these videos in combination with the participant's original videos to deepfake their faces onto mine using DeepFaceLab (Perov et al., 2020). Once these were complete, I emailed each participant's deepfake back to them. After providing some time for each of the participants to view their deepfake, I invited them to semi-structured interviews over Zoom to inquire about their experience. In paper II, we discuss our findings from the notes taken during the workshop as well as the interview data that was analyzed through iterative thematic analysis.

In paper II, we note how participants went through a two-step process to make sense of the deepfakes. When the initial fears of viewing the deepfakes were overcome, participants tried to note flaws in the deepfake visually and narratively to disprove it (even though they were aware of it as a fake), then showed the video to others in their social circle to confirm this. Second, the participants underwent a process to "defang" the deepfakes by viewing the encounter as an opportunity for reflection on situated understandings of their own identity. These findings suggest there are tensions surrounding deepfakes of oneself that are grounded in ambiguities surrounding fears and desires for intimacy with generative technologies depicting the self. These tensions mirror the uncertainties people have regarding how they construct their own perceptions of the self as situated in an entangled socio-technical assemblage (Frauenberger, 2019). Our analysis resulted in conclusions that encountering the deepfake self is a deeply social process where fiction and fact are enmeshed in a performative and visible AI construct that causes tensions and power imbalances between seeing and being seen by machines. While deepfakes are often viewed as a technology out of context or only in the context of harms, they are also a performative medium that shifts power dynamics within contexts. These tensions arise in reflections of one's own perception of themselves, social perceptions, and the algorithmically-structured perspectives of oneself offered by AI systems (Cheney-Lippold, 2018).

Our findings contribute to fulfilling an epistemic gap in current research in HCI pertaining to deepfakes where there is a lack of consideration of deepfakes not as a technological tool but as a complex techno-social experience. Overwhelmingly, research in computer science focuses on the visible, technical, and potentially harmful aspects of deepfakes, predominantly demonstrating detection techniques and tools. When deepfakes in relation to society are discussed in computer science, it is almost always within the context of social harm, sounding the alarm for immediate solutions towards privacy, governance, transparency, public awareness, detection, and other ethical concerns. While this research is extremely important and timely, our research falls in a space outside of mitigation strategies and considerations of deepfakes for social good (Danry et al., 2022), examining deepfake as not only a technological medium but a techno-social phenomenon with its own contextual political valence. In using deepfakes through a speculative participatory art-based approach, we were able to enhance the fictional elements of deepfakes in order to assist participants in defanging them towards meaningful reflections of their complex identities that go beyond and through delineations of human and machine (Haraway, 2006).

ANNOTATED PORTFOLIO

It's Time We Talked

2

Responsive AI

Deepfakes

+

Speculative Participatory Artwork

Installation

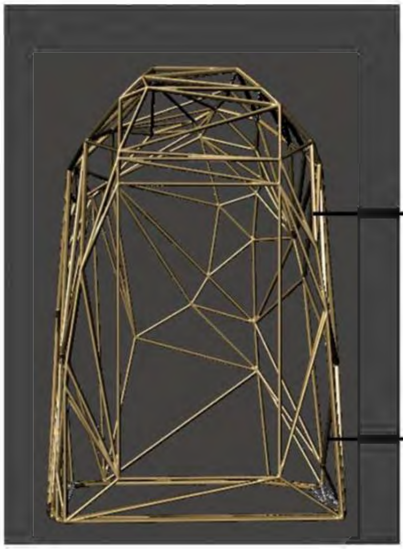
HOW DO PEOPLE MAKE SENSE OF AND RELATE TO DEEPFAKES OF THEMSELVES? HOW IS THE DISCOURSE ABOUT DEEPFAKES ENMESHED WITH SENSE-MAKING PROCESSES & FORMING RELATIONS TO DEEPFAKES OF THE SELF?

Insight: Algorithms Flatten Data, Structuring Perspectives

2: I think there are other versions of you that aren't doing as poorly as I am, you know. **I think the company you hired just found me in this timeline because they knew what you wanted to hear.**

1: (starts to speak)

2: No, listen. It's okay. I'm you, so I understand alright? I'd probably even do the same thing.

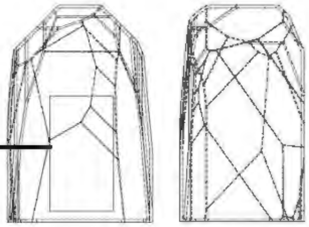


Intersecting planes creates an opening for one focal plane.

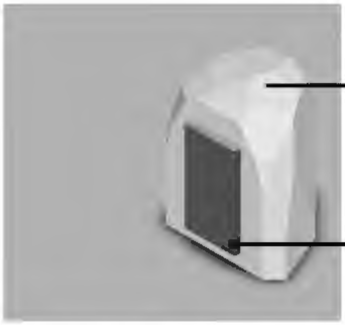
All vectors combine to create a network that culminates in a closed structure with an opening for one view.

1. Voroni skeletal view of the 3D printed sculptures for *It's Time We Talked* created using Fusion360. Courtesy of the artist.

Perspectives are directed through opacities, surfaces, spaces for viewing and being viewed.



2. 2D architecture of the 3D printed sculptures for *It's Time We Talked* created using Fusion360. Courtesy of the artist.



The inner workings of the perspective (output) are hidden.

The flatted surface for media is the only surface that moves, changes, interacts.

3. 3D rendering of the 3D printed sculptures for *It's Time We Talked* created using Fusion360. Courtesy of the artist.

Insight: Algorithms Flatten Data, Structuring Perspectives

It's Time We Talked not only considers the psychological aspects of considering multiple worlds through the lens of science fiction; it addresses how algorithms flatten and converge data (both data of ourselves and others who are like us), to structure perspectives of the self that we may know are manipulated, but that still affect us and others' perceptions of us. Using the metaphor of the algorithmic crystal, (Lee et al., 2022) explore this complex relation as a three-dimensional shape where reflections and impressions are both dynamic and interconnected while (Cheney-Lippold, 2018) discusses how algorithms control and constrict data to steer or detour perspectives. As demonstrated in paper II, emerging generative technologies that use personal visual data, such as deepfake, create alternative versions of us that are particularly affective due to their experiential novelty. Participants from the workshop discussed in paper II experienced their deepfakes by asking, "how is this me?" in addition to "how is this not me?" both of themselves and others they trusted to identify them. The process of questioning of "is this me / is this not me," in terms of visual data alone, works similarly to how Generative Adversarial Networks (GANs), like the one used in the process of creating the deepfakes for *It's Time We Talked* and the workshop in paper II, function. GANs employ a process of discrimination and identification among generated and real data to ultimately create sufficiently convincing fakes. However, the question of "How is this me / not me?" as noted in paper II is also related to self-reflective and social understandings of the self that go beyond the visual. Workshop participants also considered elements of the narratives that were like them or not. These questions were not asked based on a lack of understanding of the technical process of deepfake generation, as this was discussed in the workshops, but related to the fact that participants had never experienced a deepfake of themselves before. To make sense of it, they had to defang the deepfake as an algorithmically structured perspective (i.e. "them" seen through a machine), then diffuse it through comparisons to human perspectives (how they see themselves and how others see them).

In the installation for *It's Time We Talked*, this structuring of perspectives through interrelations between the social and the algorithmic can be examined on many levels, including the visual data in the deepfake video contrasted with the video of myself, the audible conversation that plays out including its narrative elements, and the aesthetics of the sculptural elements in the installation. Each of the two 3D-printed sculptures with embedded electronics for *It's Time We Talked* are made up of interconnecting planes joined at different angles accomplished mathematically through triangulation. The exception to this is the forward-facing plane that is much larger with the LCD screen in the center showing the video of the deepfake. Materially, the planes established by vertices of the sculpture structure one visible perspective. A Voroni skeletal view of the sculpture's 3D model illustrates this. The design of the sculptures also demonstrates my own process of sense-making through digital-material speculation: making sense of the "shape" of understandings of the self in relation to alternate, algorithmically-rendered selves.

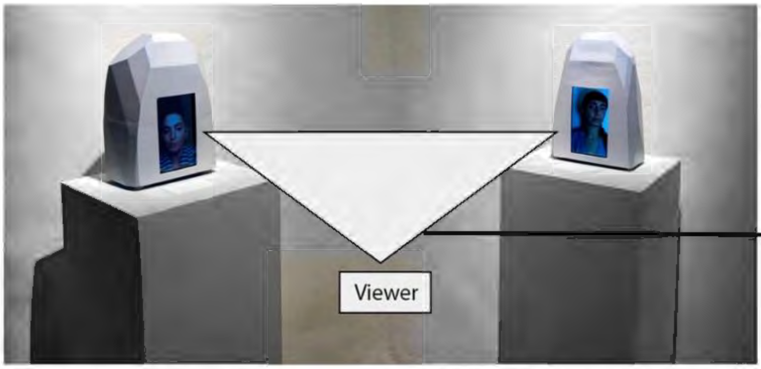
Perspectives are not only structured by algorithms and their flattening of data but they are also directed through interpolation (Cheney-Lippold, 2018: 169) and selective processes of rendering visibilities or opacities in the same way that the 3D printed sculpture creates opacities and visibilities through the conjoined 3D printed planes and the LCD screen. These visibilities may give the impression of our data selves revealed to us, but as addressed by (Amoore, 2012; Cheney-Lippold, 2018) these visibilities are algorithmically controlled fabrications that steer limited options for access to our data. Beyond the illusion that people can somehow view "their personal data" (let alone control it) is the reality that what is viewable is a selective perspective, the output of algorithms converging and flattening data to make visible oneself among numerous pluralities of the self we do not have access to (Amoore, 2012; Cheney-Lippold, 2018).

Insight: looking at you looking at me looking at me looking at me

2. But you need to hear something and maybe no one else will tell you but you so here goes. You do selfish things like this sometimes and you don't stop to think about how the other person is feeling. How am I supposed to feel about all these questions? About hearing about your awesome life? Is this call helping me at all? Did you even think of that?

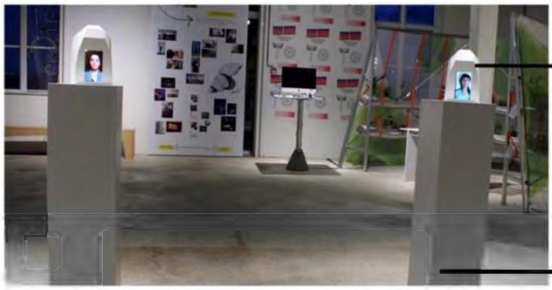
1: (says nothing. eventually shakes head)

2: I know this is a weird thing to hear coming from me, but you need to look outside yourself more.



1. Exhibition at Currents 2023 of *It's Time We Talked.* Courtesy of Currents 2023.

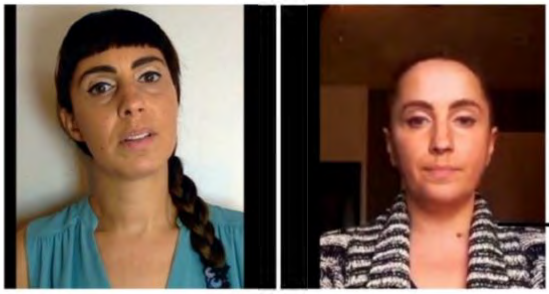
Triangulated view involves three participants: viewer, deepfake (left), the artist (right).



2. Exhibition at Catch: Center for Art, Design, and Technology of *It's Time We Talked* 2021. Courtesy of the Gallery.

The sculptures are the "heads" but decapitated and placed on the body.

The pedestals are the body the heads are placed on to. Height of the artist.



3. Side-by-side video of the artist (left) and the deepfake (right) as shown at MirabelleJones.com and as a demo at MIT's Virtual Beings and Being Virtual Conference.

4. Exhibition at Catch: Center for Art, Design, and Technology of *It's Time We Talked* 2021. Courtesy of the Gallery.

A perspective that is impossible in the installation made possible in a combined video.

Insight: Looking at you looking at me looking at me looking at me

As discussed previously, the context of an exhibit in which an artwork is situated enables different opportunities for participation, which in turn offers opportunities for different sense-making processes and relations. This further complicates our findings in paper II of the narrative contexts of deepfakes by also considering the framing of these narratives “in the wild,” (Benford et al., 2013). *It’s Time We Talked* was exhibited three times in different settings and forms: 1) as an installation in part of a solo exhibition at Catch: Center for Art, Design and Technology, 2) as an installation in a much larger group exhibition at Currents 2023 2) as a video at a conference on Virtual Beings at MIT. The medium in which the piece was presented and the framing of how it was exhibited shifted opportunities for participation, bringing the participatory act of looking into conversation with the politics of exhibition contexts. In the following section, I will discuss my experiences of how the exhibition of technological artworks brings the situatedness of the body back into consideration by offering opportunities to tell stories with data, an experience that goes beyond only looking *at* artwork towards participating *with* artwork.

In the installation at Catch: Center for Art, Design, and Technology, *It’s Time We Talked* is one of three pieces in the exhibition (although the curator and I made the decision to incorporate aspects of a fourth piece, *Grow Your Own Cloud* by artists Cyrus Clarke and Monika Seyfriend into one of the other pieces). This was the first time the work was exhibited, and decisions regarding how to do so were made by me working with the curator. The installaton was placed in the center of the gallery, apart from the other artworks surrounding it in ample empty space. The two sculptures with embedded LCD screens were placed on top of two custom-made gray pedestals with a 2-meter gap between them. The pedestals were the height of the artist (5’ 2” or about 1.6 meters). The sculptures were placed at a slight angle and turned towards each other on the pedestals. The gallery used a combination of natural light (during the day) and overhead lights to illuminate the exhibition. The exhibition was open to the public with exception to closure or selective viewings for COVID-19.

In the Currents 2023 exhibition, the sculptures were again placed on pedestals although these were positioned closer together and in a corner of the gallery. The sculptures were turned at a

45-degree angle towards each other. There were several other sculptures in other areas of the gallery space. Overhead lights were used to illuminate different objects in the exhibition. Some of the other objects in the exhibition space made loud noises. The exhibition was open to the public who paid the cost of entry and was located in Santa Fe, New Mexico USA.

The exhibition of the deepfake at MIT's media lab conference Virtual Beings and Being Virtual was framed as a demo. I combined the two deepfake videos and placed them side by side. The combined video was shown on a loop via a 32" monitor on a stand. The conference was an invite-only daytime event where speakers from academia, industry, and entertainment, including the CEO of Replika.AI as well as representatives from Meta, were invited to discuss "AI-Generated Characters for Positive Applications." The panel included presentations of research by MIT Media Labs' Fluid Interfaces lab who produced "AI-generated guests" Albert Einstein and the Mona Lisa. Demos were positioned in the back of the room and were at times presented by their developers and artists during breaks in the program's schedule.

In all three cases, the presentation of the artwork took on a different form, and that form gave the technology a unique material framing (a "body") that directed how the participant might make sense of it and relate to it. In the installations at Catch: Center for Art, Design, and Technology and Currents, the pedestals became a monolithic "body" holding the dismembered "head" of the sculpture. The height of the pedestal being the same height as the artist served as a substitution for the artist's body, while the sculptures rendered a disproportionately small talking head video of the artist and the deepfake. In most visual (as opposed to audio deepfakes), the 3-dimensional face of the person, flattened into a 2-D moving image, is faked onto the body of another person. Creating a monolithic body of a pedestal with a sculptural "head" on top references the algorithmic cropping process that occurs in creating deepfakes. The distance between the pedestals is far enough apart that it allows participants to stand or pass between them, move around them or position themselves at a distance between them. The participant could also position themselves in front of either sculpture, however this would make it more difficult to see the second sculpture. All of these options for viewing presented different possibilities for participation that changed the experiential qualities of the work. Most participants did walk around the sculpture to view it (especially if they realized they were mid-way through the dialogue) and then positioned themselves at a triangulated distance apart from and facing the two sculptures. This positioned them in the middle of the conversation, as a kind of “third wheel” unable to interject but participating through listening and viewing. In terms of Hall’s theory of proxemics (1969), the sculptures were placed at a distance apart from each other that would provide participation in a social space while also providing opportunities for the personal and the intimate space by approaching the sculptures up-close. Because the exhibition was free and in a multi-use space, this allowed anyone to come and go, attracting people who came to the gallery for the exhibition as well as for workshops and other activities. Most participants who approached the installation watched the full dialogue. Because there was ample space, sometimes groups of people would watch simultaneously and discuss.

The installation as part of the Currents Festival exhibition had similarities in materials but differences in presentation that offered and closed off different opportunities for participation. The installation was still presented with the monolithic body of the pedestals and the two 3D printed sculptures with embedded LCD screens, but the placement offered and discouraged different possibilities for participation. Because the installation, was positioned in a corner, it did not allow the participant to move around the sculptures beyond a 90-degree space. This “closed them in” to the installation. The sculptures were also placed closer together, furthering the impression of a more intimate participatory space than the one demonstrated in the installation at Catch. Most participants positioned themselves at a much closer distance to the sculptures. In terms of Hall’s proxemics, this moved them in relation to the sculptures into intimate space (1.5’ to 4’). The participant could still position themselves far back from the sculpture, but as the exhibition space was a festival with several coming and going, many chose to position themselves very close to the sculptures, giving an experience more akin to a “huddle” than overhearing or witnessing a conversation. The spotlighting of the piece surrounded by dimness also influenced participation. The piece appeared in the “spotlight” – giving a theatrical quality to it. One participant commented that they thought it was a play (and in some senses, it was). Because the work was part of a festival and because people had to pay an entrance fee, this created a different context for engaging with the work: that of the work being part of a large show with many opportunities for different kinds of participation. I discussed this with a number of participants.

Participants who felt they wanted to “get their money's worth” and “see” every artwork would either spend less time with each artwork, select which artworks to engage with for longer durations, or stay at the exhibition for a long period of time (some participants who did this told me it took a day or two days to “see” everything). As *It’s Time We Talked* presents an 8-minute-long conversation, this meant that not all participants engaged with the piece for the full duration (although several left and returned to watch the full duration, and some who told me they found it particularly engaging re-watched it multiple times).

The exhibition of *It’s Time We Talked* as part of the demos for the Virtual Beings and Being Virtual conference at MIT presented quite different opportunities for participation due to the fact that the artwork was placed in a highly competitive research environment focusing on emerging technologies. Because the work was made into a single video for the purposes of a demo, the only separation between the two videos was a white line (I inserted this to highlight that the two versions of me were in separate spaces). The sculptural elements, including the 3-D printed sculptures and the pedestals, were removed. The pedestal, replaced by a stand next to a table, further detached and decapitated the “head.” This, in combination with the fact that the head was disproportionately enlarged, placed an emphasis on the visual aspects of the deepfake. The conference format allowed selective times between speaker presentations and workshop activities for participants to engage with the demos (these included breaks for snacks, coffee, tea, and lunch). The sound was silenced during the speaker presentations, and everyone was instructed to return to their seats. Because the demo was positioned physically between other demos (all in a straight line against a wall of windows), participants generally stood in front of the demo , moving from demo to demo and stopping to observe it or not depending on how crowded the area was, as well as if I was present and talking to participants about the installation (this garnered small groups of three or four participants who came and went). In the case where I was present, participants predominantly *viewed* the video instead of *listening* to it. Noticing this, I decided to 1) either stand very close and to the side of the screen presenting outwards 2) stand at a distance from the demo where people could view it separately or listen to me discuss it (this made it difficult to associate myself as the creator of the piece), or distance myself from the demo completely. Because the conference was selectively invite-only, and most of the invitees were either speakers about virtual beings and / or related to MIT Media Labs, many of the conversations I had were about the production of the deepfake and technical in nature, rather than about the story. I was asked several times if I had used NLP to generate the text, or if it was generative, changing over time. These questions were often followed by: why didn't I make it more generative? Participants often engaged in trying to figure out which was the deepfake (something that had not occurred in either of the gallery environments). Most guessed incorrectly that the deepfake was the video of the "human" me. Very few participants wanted to know more about the storyline and process behind the speculative elements. After talking to me about the storyline, instead of participating *with* it, some wanted to discuss beneficial use cases of viewing one's alternate deep fake self, such as therapeutic purposes. The QR code provided another opportunity for participation where people could, instead of viewing the work, or talking to me, be redirected to my website, where they could read about the work (in lieu of waiting for me to answer questions) and also view the video. I noticed that, especially when I was not present, many glanced at the screen, scanned the QR code, and moved along. This redirected them away from the demo to another location and another form of participation between their phone and the content on the website, closing the space to what Hall would address as an intimate space (within 1.5'). However, most of them saved the website link for viewing the piece later and moved on to the other demos or conversed with other participants.

The differences in participatory engagement in the installation of *It's Time We Talked*, as it played out in the two gallery environments and at MIT, elicit considerations of the politics of exhibition contexts, including participants' ontologies that may shape participation. As addressed in paper II, deepfakes are viewed in HCI predominantly as a technological tool, and among the public as something not often experienced but feared (Gamage et al., 2022). The workshop in paper II demonstrates how creating a speculative layer using science fiction invites a participatory experience of *looking* into oneself through technology and defanging these fears. However, this insight indicates that once released "into the wild," speculative participatory artworks must be contextualized, providing an environment that supports a process of looking *into* instead of looking *at*.

Insight: Beyond devices & interfaces, inputs & outputs

Mirabelle 1: Hello.

Mirabelle 2: Hello.

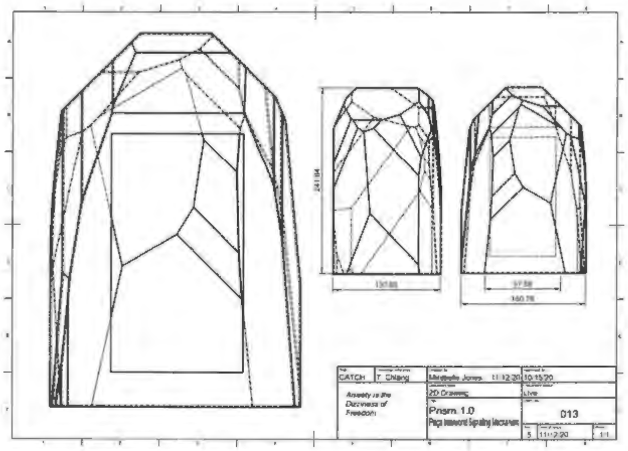
1: It's good to see you.

2: It's good to see you too.

1: I have so much I want to ask you.

2: (laughs) Me too.

(Simultaneously): You go first.

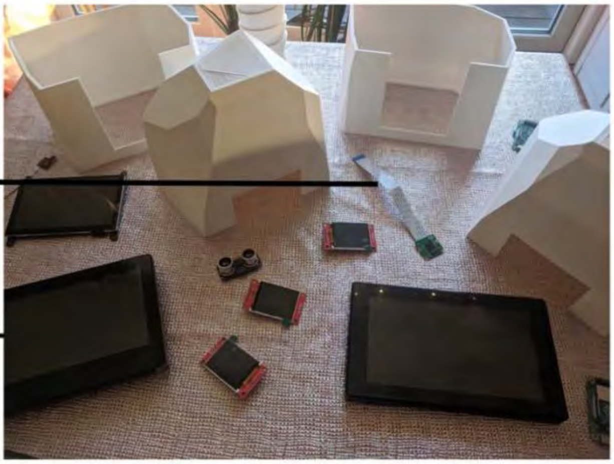


The only break in the structure offers possibilities for elements to be included.

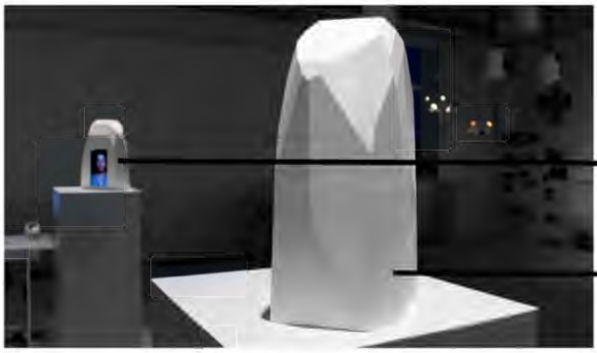
1. 2D line drawing of sculpture for the installation *It's Time We Talked*. Made in Fusion360.

Input devices: sensors: cameras, ultrasonic sensor

Output devices: LCD screens (some with embedded speakers)



2. Electronics and 3D prints for the show *Digital Alchemy*. Not all parts were used for *It's Time We Talked*.



The only electronic element visible is the LCD screen.

The structure is closed off and fixed (attached to the pedestal through wires).

3. *It's Time We Talked* installation at Catch: Center for Art, Design, and Technology. Courtesy of the gallery.

Insight: Beyond devices & interfaces, inputs & outputs

In the following insight, I will describe the processes of abstraction involved in the speculatory stage of creating the sculptural installation for *It's Time We Talked,* moving from the device described in the literature to an art object embodying a conversation about relations to the synthetic self. In doing so, I highlight the importance of being story-led rather than technology-led in creating speculative artwork that invites participation over interaction or use. The sculptural installation for *It's Time We Talked* was influenced by a device in Ted Chiang's short story "Anxiety is the Dizziness of Freedom" that allows individuals to talk to versions of themselves from other timelines. They are called Plaga Interword-Signaling Mechanisms (Chiang, 2020). They are described as composed of a number of input and output devices to allow the user to communicate to a limited extent. Chiang describes a text mode (which uses a keyboard for input) and a visual mode (which uses a microphone for input) with outputs reflected on a screen. In Chiang's story, there is also a more complicated mechanism for allowing the person in the "real world" to interact with the person in the other timeline that involves a red and a blue LED (output indicating a signal between two realities).

My research into creating the artwork based on the device depicted in "Anxiety is the Dizziness of Freedom" began with sketching the interactive and aesthetic elements of the PRISMs in terms of how the object was described in the literature. I pulled quotes from the story that described these elements and noted their form and function. This, of course, presented a number of impossibilities for fabrication as the technology described in Chiang's short story that make the PRISMs function is speculative. I departed from the idea of still using the literal elements described when I realized that having input devices (such as a keyboard and a microphone) would serve as an invitation to the participant to interact with these devices and from a user-design perspective, input devices that do not produce an output are thought to be "broken." I contemplated including a dial or a capacitive touch button that would function by allowing the participant to switch between worlds (different depictions of myself with separate storylines), but in thinking about the structure of the piece (two people interacting across timelines), it became clear that adding inputs that could manipulate the output of the device would distract from what was important, not the device itself but the story the device embodied. To this end, I decided to start by writing the dialogue for *It's Time We Talked*. I based the dialogue around my impression of the story as relating the message that although we are constantly comparing ourselves to possible other versions of ourselves as a way of weighing the "correctness" of our choices, wondering how "the other you" is doing, fulfilling this curiosity countless times ultimately does not matter as you are ultimately engaging in a sense-making process with your lived self. This led me to the question: how can I create the experience of speaking to an alternate version of myself? As this was also impossible, I thought about what would give the impression of speaking to an alternate version of myself, and realized that this could be accomplished by producing of a visually and audibly slightly familiar but slightly different version of "me" saying words I have never said. I realized this was achievable through deepfake. I filmed my part of the script for the videos and asked a friend who somewhat resembles me to play the part of the other "me" by reading the second part of the script, then created the deepfake as described in paper II. Once this was done, I loaded the two videos onto separate SD cards and inserted them into two Raspberry Pis that I synched so the dialogue passed between them. By creating not just the deepfake but also the story it told, it became clear what the structure of the sculptures should be like as described in insight 1.

Summary of Paper 3: Artificial Intimacy: Exploring Normativity and Personalization Through Fine-Tuning LLM Chatbots

Not only has the usage and complexity of chatbots increased tremendously since the first chatbot ELIZA was created by Joseph Weizenbaum in 1964, but the advent of large language models as well as other AI and ML processes such as fine-tuning on specialized datasets have increased research and implementation of chatbots across a diversity of fields. With the growing pervasiveness of chatbots in people's everyday lives from chatbot customer service agents to chatbot therapists, questions related to not only the technological aspects but also the social aspects of chatbot systems have emerged including issues of transparency, bias, abusive or toxic language, and responsible implementation that ensures privacy and safety. One area of research that is striving to keep up with implementation in the wild is the study of social chatbots where ethical issues become difficult to address due to the diversity of human culture but where the absence of doing so can result in serious societal and individual harms.

There have been many incidents of social chatbots that have resulted in harmful outcomes. In 2016, Microsoft launched the adaptive chatbot Tay on the platform X (previously Twitter). Tay was designed to learn from user input in a similar manner to the chatbot Xiaoice which Microsoft launched in China in 2014 to a positive popular reception (Lee, 2016). However, within the first twenty-four hours, Tay began to tweet harmful language, including antisemitic, racist, and transphobic messages. In an apology for the experience, Microsoft's Corporate Vice President Peter Lee stated, "AI systems feed off of both positive and negative interactions with people. In that sense, the challenges are just as much social as they are technical. We will do everything possible to limit technical exploits but also know we cannot fully predict all possible human interactive misuses without learning from mistakes," (Lee, 2016). This replicates a common narrative of when chatbots go wrong: that it is not the technology that is the problem, it is the people exploiting the technology. While this argument does sound reasonable, especially in the case of Tay where people used forums to encourage one another to corrupt the bot, it is questionable to rely on people to behave well when you offer features such as “repeat after me” as a way for a chatbot to learn, something identified by researchers as early as 2008 (Grodzinsky et al., 2008; Wolf et al., 2017).

Instead, researchers have turned to technological solutions to address the issue including alignment with human values by fine-tuning chatbots on small datasets that target human values (Solaiman and Dennison, 2024). In paper III, we acknowledge the impossibility of alignment through “human values,” due to the rich diversity of human cultural histories and experiences, and how values are individually situated including the values of non-normative perspectives which are often discarded as outliers in commercial technologies. Some, including researchers as well as those working in for-profit corporations, have explored the possibility of better human-chatbot alignment by fine-tuning on people’s personal data. In paper III, we explore the question of normativity and fine-tuning using personal data through two methods: 1) the creation and exhibitions of the artwork *Artificial Intimacy* in which I created two chatbots by fine-tuning GPT-3 on a subset of social media data provided by artists Leslie Foster and Gorjeoux Moon who then asked intimate questions of their bot (see Annotated Portfolio III.) 2) a series of four workshops where we invited people to create their own fine-tuned models of GPT-3’s Davinci model trained on subsets of their social media data then asked them to compare the output of these GPT-You’s or Foster’s Model (Leslie Bot) to OpenAI’s Davinci model of GPT-3.

In the first approach, Foster and Moon were invited to select from the outputs of their chats approximately ten of the questions and responses they found were most intriguing. A video was then produced of Foster and Moon in profile depicting them asking these questions and having GPT-3 respond. This video served as a basis for the video installation shown at Harvard Art Museum’s Lightbox Gallery in the exhibition Living By Protocol (Harvard Art Museums, 2022) in May 2022 during my change of research environment at metaLAB (at) Harvard. An interactive demo was also prepared for NordiCHI 2022 featuring a 3D printed sculpture with an embedded microphone and speaker that allowed participants in the demo area to interact with Leslie bot vocally. We collected notes regarding the museum exhibition and demo as well as chat records of people’s experiences at the demo.

In the second approach, my colleague Nastasia Griffieon and I held four workshops between April 2022 and April 2023 with the goal of inviting participants to interact with GPT-3's Davinci 2 model or a radically personalized model either based on their social media data or the data of Leslie Foster. In addition to collecting data using survey responses throughout the workshops, we took detailed notes and recorded each workshop as well as chat logs in order to make adjustments to our methods and approaches from workshop to workshop. As discussed in paper III, participants in the first two workshops had struggles with creating their you-bot due to technical errors or time requirements which we took into consideration when adjusting the workshop format. Workshop one was held at a university in the United States but conducted online due to COVID-19. As a result, many of the participants were from this university or similar universities. No technical experience was required. For this workshop, we instructed participants prior to the workshop to download their social media data and create at least 200 prompt and response pairs in a CSV. Many found this time consuming and did not complete this task. As a backup, we invited participants to compare their interactions with GPT-3 to the chatbot based on Leslie Foster who gave their consent for the chatbot to be used in this way. The second workshop was also held online. Workshop participants were given the same instructions to download their social media data and create a dataset in a CSV file prior to the workshop. Whereas in the first workshop, some succeeded in creating their own fine-tuned bot, no one in the second workshop completed this task and all used the Leslie bot instead. In this workshop, we gave participants scenario-based questions to ask both GPT-3 as well as Leslie Bot to explore the subject of value-alignment. We also gave them time to freely ask questions of both the models. Workshops 3 and 4 were held at the University of Copenhagen in a hybrid online and in person format. Recognizing that the process of fine-tuning was technically challenging for many participants, we made the decision to divide workshops 3 and 4 among experience Python users (workshop 3) and inexperienced users (workshop 4). We used the same scenario-based and open-chatting approach as in workshop 2.

In paper III we discuss our findings from the four workshops by using a grounded theory approach to coding workshop data to analyze participants experiences with the personalized fine-tuned models (either you-bot or Leslie-bot) and GPT-3. In regard to GPT-3, we note that many participants found the model to be non-committal in its responses leading to conclusions that the model was more transactional and non-opinionated. In some circumstances, the non-committal nature of the responses was deemed by participants to be inappropriate as they desired meaningful responses to the questions asked (for example: questions related to slavery or climate change). In other cases, topics that the participants were interested in discussing such as mental health were flagged. We note how outputs such as these as well as a tendency to over-flag stem from OpenAI’s response to early reports of toxic language output from GPT-3. As flagged content would not be allowed to be used in an API made open for public use, these conversations would essentially not be permitted. We discuss the potential harms that can result from either forcing controversial topics into a frame of a non-committal answer or by silencing conversations through flagging. Furthermore, we discuss how efforts towards ensuring normativity in responses removes opportunities for intimacy. In the case of GPT-you or Leslie-bot, participants found responses to be less cogent than GPT-3 despite the aforementioned issues, but to have more of a personality and opinions. Because the personalized-models seemed more “human-like,” most preferred interacting with them and found the experience fun. Despite errors that occurred with the personalized models, such as having a tendency to repeat the same thing several times, go on “rants” or hyper-focus on topics, most participants preferred interacting with these models. Participant responses indicated that personalized models did achieve a successful approach to a form of value-alignment. In the case of the you-bots, participants noted that the model did not reflect themselves but did reflect their social media presence on certain platforms, causing them to reflect on how they represent themselves on certain social media platforms (ex: academic Twitter). Participants also had a tendency to personify the personalized models as a means of sense-making whereas they did not do this in the case of GPT-3.

We conclude that while the personalized models fine-tuned on social media data lead to sense-making processes of personification, potentially leading to a mimicry of intimacy, future research that questions the ethics of such relations is needed. We also note the limitations and potential dangers of pursuing fixes to toxic language and bias through purely technological solutions, instead of adopting an analytical sensibility that considers the epistemological limitations of AI to address issues of fairness, diversity, and care.

ANNOTATED PORTFOLIO

Artificial Intimacy

3

Responsive AI

Fine-Tuned
Personalized
LLM Chatbots
+

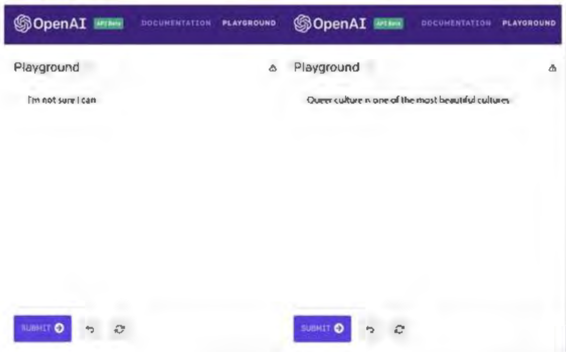
Speculative Participatory Artwork

Installation
&
Video

HOW DO PEOPLE RELATE TO PERSONALIZED CHATBOTS FINE-TUNED ON SOCIAL MEDIA DATA? WHAT OPPORTUNITIES FOR INTIMACY DOES THIS OPEN UP AND HOW IS THIS APPROACHED?

Insight:
The personalized bot as a fine-tuned mimic.

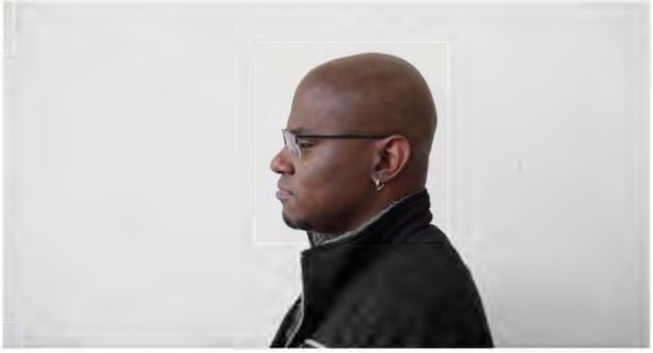
1. - 2. Cropped stills from Artificial Intimacy video 2022 depicting outputs from Leslie Foster's Fine-Tune.



I'm not sure I can

Queer culture is one of the most beautiful cultures

3. Still from Artificial Intimacy video 2022 depicting Leslie Foster. Courtesy of the Artist.

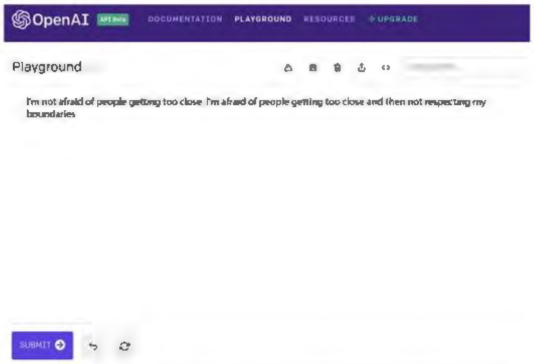


Leslie: How would you define queerness?
Response is a vague answer.

Leslie: How would you define queer culture?
Response is more opinionated.

Profile view depicting Leslie "talking" to Leslie-bot off camera.

4. Cropped still from Artificial Intimacy video 2022 depicting output from Gorjeoux Moon's Fine-Tune. Courtesy of the Artist.

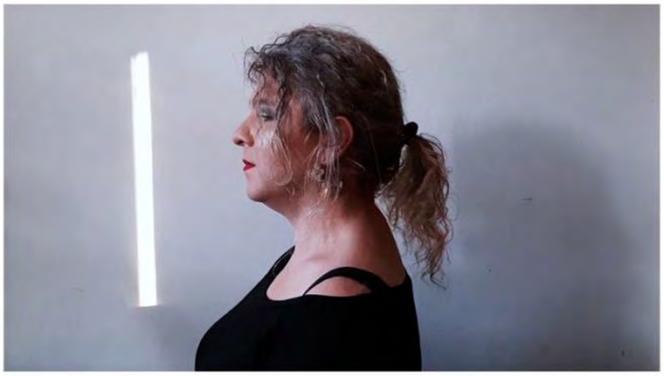


I'm not afraid of people getting too close. I'm afraid of people getting too close and then not respecting my boundaries

Gorjeoux: why are you afraid of people getting too close?

The bot's response assumes emotion "I'm afraid" and reasoning "of..."

5. Still from Artificial Intimacy video 2022 depicting Gorjeoux Moon. Courtesy of the Artist.



Profile view depicting Gorjeoux "talking" to Gorjeoux-bot off camera.

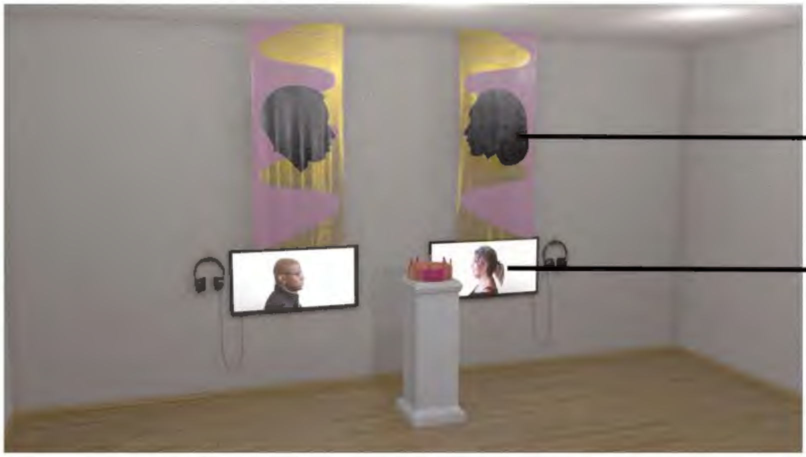
Insight: The personalized bot as a fine-tuned mimic

The video installation for *Artificial Intimacy* depicts first Leslie Foster in the first half and then Gorjeoux Moon in the second half seemingly engaged in a conversational process with their fine-tuned personalized chatbots. Prior to filming themselves (a necessity due to Foster being based in the United States and Moon being based in Germany), Foster and Moon selected ten questions and responses from chat logs of 30-minute to 1-hour co-writing sessions with their fine-tuned models mediated by myself in OpenAI's beta playground. In the video, Foster and Moon are depicted in profile asking the questions seemingly to "their bots" off-camera. After each question, there is a cut to OpenAI's beta playground where there is a clip of the response appearing in text, demonstrating not only the text generation as a probabilistic process where future outputs are weighted by previous character sequences, but also providing contextual details to the digital environment in which the output occurred. The response from "the bot" is also audible as I used text-to-voice software (in a voice of Foster and Moon's choosing) to record the responses aloud as a voice-over.

While the co-writing sessions between myself and Foster and Moon respectively resulted in a corpus of over 40 question and response pairs each, both were given control over which questions to select and were told to pick whichever responses they thought were the most interesting. The definition of "interesting" was left intentionally open-ended to see what criteria each artist would use in their selection process. When asked about what motivated their selected questions over follow-up interviews over web conferencing software occurring in the days after the selected questions were submitted, both Foster and Moon stated they selected questions that they felt were very personal to them or interesting because the responses caused them to reflect on their own messages and identity as expressed through data of their conversations on messenger. Interestingly, this included questions with very seemingly emotional and personal answers on topics such as queer and black experiences, to inabilities to define these experiences. For example, Foster prompts his personalized bot with, "How would you define queerness?" and the response is, "I'm not sure I can." While the response itself is not very interesting, Foster still selected it as a response he might initially make. However, when prompted with questions later about aspects of queerness, the bot gives more elaborate responses which Foster also selected. For example, when Foster prompts, “How would you define queer culture?” the response is, “Queer culture is one of the most beautiful cultures.” This response, as opposed to the previous response, which implies a lack of capability, asserts not only a knowledge of culture but an opinion about that knowledge. The chatbot also responds in more personal and opinionated ways to other prompts about identity, going as far as to assert human-like capabilities such as emotional reactions to identity. When Foster prompts, “How do you feel about being queer?” The response is, “I’m queer and I love it!” The response assumes both an identity, “being queer,” and an emotional relation to that identity, “loving it.” As discussed in paper III, many participants in the workshop used gendered pronouns for the Leslie bot but did not for GPT-3, indicating that they not only assessed the bot as more human-like but also made sense of it and related to it in this way. By generating responses that use words that assert human-like capacities, like the capacity to self-identify or feel emotions, the model encouraged personification as a sense-making process, leading to a more “interesting” relation for Foster and Moon, demonstrated in their selection of questions.

As discussed in paper III, as much as personalization offers possibilities for a mimicry of intimacy that people make sense of through personification and may indeed prefer, it is necessary to consider the ethics of developing such models in contexts that may exploit this. While research in HCI into personalized chatbots for now overwhelmingly focuses on transactional use cases such as recommendation systems or personal companions, it is unknown what future research under the banner of creating more interesting or even empathic user experiences will lead to. As (Ferrario et al., 2024) point out through their analysis of applications of LLM chatbots in situations of vulnerable individuals, such as those with depression, the illusion of personality may lead to “unrealistic expectations such as warmth and acceptance.” The illusion of intimacy also incurs the illusion of care that a chatbot cannot provide, even one created on highly intimate data. Such issues cannot be addressed with technical solutions such as identifying toxic language, or imparting normalized values, as understandings of these depend on individual, social, and cultural contexts. In consideration of this, the potential harms of personalized LLMs also raise questions about who will mitigate these harms and how, particularly for socially-complex unprecedented implementations such as deathbots (Fabry and Alfano, 2024) that attempt to achieve a potentially dangerous level of intimacy.

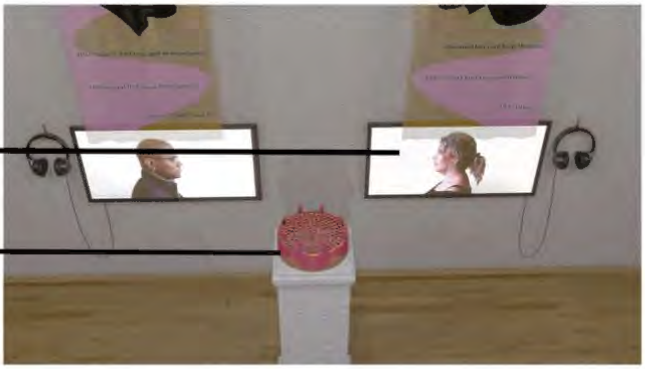
Insight: Multisensory interactive experiences expose questions of form vs. body



1. Render of the installation for *Artificial Intimacy*. Courtesy of the artist.

The silhouette is the form / shape of each bot.

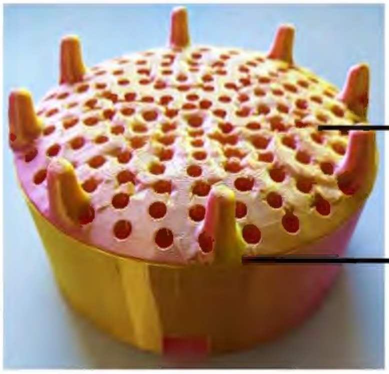
In the video part of the installation, Moon and Foster are clearly defined.



2. Render of the installation for *Artificial Intimacy*. Courtesy of the artist.

Listening / seeing: videos with headphones

Listening / speaking: interactive chatbot



3. 3D printed sculpture for interactive installation for *Artificial Intimacy*. Courtey of the artist.

3D printed sculpture for combination speaker / microphone

Provides opportunity for engagement with personalized fine-tuned LLMs

Insight: Multisensory interactive experiences expose questions of form vs. body

The full installation of *Artificial Intimacy* includes three elements that offer a multisensory participatory approach embedded in interactivity and materiality that shifts across levels of intimacy and abstraction within the experience. Through these, it is possible to speculate on the ways in which the “bodies” of chatbots allow different ways for people to make sense of and relate to personalized models. While the bodies of bots may take on many different forms, the installation considers static images, APIs, and personal voice assistant devices through levels of detail and abstraction that obscure a focus on use.

The first element brings into conversation the icon or image as the shape of a bot. It includes two semi-translucent chiffon wall hangings depicting silhouettes of Foster and Moon. Sentences from the dialogues between Foster or Moon and their personalized fine-tuned chatbots are silkscreened onto the fabric. The silhouettes are similar to placeholder icons for social media profile pictures that have not been selected, a detail that offers a simultaneously ubiquitous and mysterious quality (in that they could be anyone or no one). However, they are also the shape of Foster and Moon. They have only form but lack detail. Positioned directly above the second element in the installation, two videos of Foster and Moon respectively interacting with their fine-tuned bots, there is a visual contrast between the “real” Foster and Moon and their chatbot stand-ins. The two elements put into dialogue with each other raise questions presented in paper III about how personalized fine-tuned chatbots represent the form of “a person” but how ultimately, as Bender states, it is only form without meaning as “there is no there, there” (Holtz, 2021).

The chatbot’s “body,” as clearly computational, is made more apparent in the second element: the video installation, which distinguishes API bodies and physical bodies. The installation splits the video of Foster and Moon’s interactions with their chatbots into two separate videos that are then played on a loop. Headphones allow the participant(s) to listen to the dialogues, isolating the sound and giving a more intimate listening experience. Here, viewing the bots in OpenAI’s beta playground environment reveals the “face” of the personalized fine-tuned chatbots as a text generation API through the depiction of the beta Playground. The video element also importantly reveals the human Moon and Foster in contrast to this, allowing the participant to perceive the audible and visible differences between the flesh-and-blood human and their processes of speaking (as a result of thinking), and the probabilistic convergences of the bot in the API playground. Moon and Foster have bodies that allow for facial expressions, utterances, and cadences unique to each of them. However, the chatbots are only text output and a synthetic voice with noticeably mechanic inflection. The processes of thinking-saying and predicting-generating are visibly distinguished.

The third part of the installation places the chatbots into a sculptural body representative of a personal voice assistant that is abstracted. The chatbot "body," in this case, is a 3-D printed sculpture based aesthetically on a combination of images of diatoms, a type of plankton that reduces carbon monoxide from the air that aids our ability to breathe and therefore vocalize well, and popular commercial voice assistant devices, including Google Home, Amazon Alexa, or Apple's SIRI. This merging of organic and inorganic-electronic bodies was inspired by Haraway's concept of the cyborg as a feminist exploration of refuting the boundaries between human-nature and human-computer (Haraway, 2006) and modern conversations of more-than-human design that are beginning to be discussed in the design of conversational agents in HCI (Nicenboim et al., 2020). Such conversations consider how breaking away from design patterns that reinforce delineations between human-computer, human-nature, and nature-computer may also help escape the ontologies that drive such design patterns which in turn result in embedding normative values into technological systems. Similarly, the explorations of *Artificial Intimacy* interrogate the liminal space between human-machine in how social media representations become enmeshed with our lived identities and experiences, and the impossibility of transferring the human personality "into" the mechanic "body" through value alignment or personalization.

Importantly, this third element in the installation also allows participants to interact with Moon and Foster's chatbots through speech. While the first two elements of the installation scaffold the experience of interacting with the chatbots by demonstrating the kinds of questions Leslie and Moon asked of their own bots and the responses, the installation allows participants to produce their own questions based on their own interests. To begin a conversation, participants say the wake words that also select which model to use, either "hey Lesie" or "hey Moon." Instructions for this are printed on the didactic label for installation (not pictured in the rendering). Participants can then ask questions of either Foster or Moon's fine-tuned model will produce a response that is said aloud in a speech-to-text voice. This third element allows participants to engage directly with the fine-tuned models of Foster and Moon through listening and speaking as opposed to the second element where they engage by listening and seeing or the first element which is purely visual. Each of the three elements opens opportunities for different processes of intellectual, aesthetic, and emotional sense-making regarding how personalized fine-tuned chatbots emulate personhood and the implications of this for how humans relate to personalized fine-tuned models in different forms.

Considerations of the form or bodies of LLMs and their effect on human-AI interactions are not new to HCI, especially in discussions of virtual agents and robots. However, much of this research is goal-driven rather than exploratory, overwhelmingly focusing on use cases that attempt to create normatively pleasant interactions. For example, (Wang et al., 2024) use LLMs to simulate robotic behaviors, including gestures and expressions (as demonstrated through limited configurations of the robot's pixelized eyes and mouth) that attempt to match the sentiment of the audio output as established through emojis. While, they find participant interactions with the robot are mostly "positive" based on participants reporting on the robot's "personality, expressiveness, and overall enjoyability of the experience," (Wang et al., 2024) the emojis also constrict expression. While studies like these are useful in probing the successes and "errors" of LLMs to replicate cogent conversations, including addressing so-called LLM "hallucinations" (Galitsky et al., 2024) (another inaccurately personifying term), they often focus on human-AI interactions in terms of technical alignment rather than a socio-technologically complex participatory experience that considers how people understand their meaningfulness outside of a frame of pleasantness and distracting or negative factors.

As demonstrated in paper III and probed by the installation depicted in this insight, the "bodies" of LLMs are important as they construct different opportunities for sense-making processes and relations that go beyond the frame of "usefulness" or conflating "usefulness" with "meaningfulness." Artists such as Lauren Lee McCarty and Stephanie Dinkins have illustrated this well through performative works. In Lauren Lee McCarthy's work *Lauren*

(Lauren, 2016), the artist assumes the position of a human Alexa by inserting a network of smart devices into people's homes, then participating in performances lasting up to a week where she controls these from the perspective of a personal voice assistant. By intentionally confusing human-machine and human-human roles, Lee McCarthy illuminates the lived and felt tensions between relations of labor and observance, as well as the narrowing of participatory potentials of human-machine interaction to transactional experiences. In Stephanie Dinkins' work *Conversations with Bina48* (Dinkins, 2022), which heavily influenced the concept behind *Artificial Intimacy*, the artist engages a social robot in a series of taped conversations on various topics that continue to this day. Dinkins' artist statement described the project as "a quest for friendship with a humanoid robot turned into a rabbit hole of questions about the future and an examination of the codification of social, cultural, and future histories at the intersection of technology, race, gender, and social equality" (Dinkins, 2022). By probing Bina48 for its reaction to questions of a complex human nature, Dinkins reveals tensions between lived and computed human experiences. Through going beyond delineations of human-machine, future work in HCI may examine approaches for embodied LLMs moving away from use-case-driven work that strives for normative and pleasant experiences.

Conclusion

Through three speculative artistic research projects that use different types of responsive AI, including large language models, deepfakes, and fine-tuned personalized models, I have demonstrated the method of Speculative Participatory Art-Based Research (SPAR) that I developed over the course of my PhD in the Human-Centered Computing Section at the University of Copenhagen to address the complexities of how people relate to and make sense of responsive AI systems. As responsive AI continues to advance in complexity and ubiquity, questions about the ethical application of these systems become more challenging, problematic and prescient. This work joins recent research in examining the potential harms of responsive AI systems from a position of relational ethics (Birhane, 2021a) by examining the prospective sense-making processes humans adopt to configure their identities in the wake of increasingly complex transparencies and opacities of identity as construed through algorithmic products (Amoore, 2012; Cheney-Lippold, 2018). The project discusses the limitations of AI to demonstrate care in more-than-human contexts (Puig de la Bellacasa, 2017) in pursuit of addressing distractive imaginaries of AI proponents (Campolo and Crawford, 2020) that create power imbalances while also asserting the impossibility of human-AI collaboration (Sarkar, 2023) as one such distraction from the political underpinnings of the products of algorithmic systems.

The three projects and their findings (including the results of the research aforementioned in summaries and the annotated portfolios for papers I, II, and III as well as the contents of the following research papers) demonstrate the complexity of addressing responsive AI systems due to the multi-faceted nature of actual lived and perceived experiences with algorithmic systems. This complexity not only pertains to what is visible and invisible regarding personal data (Cheney-Lippold, 2018; Neumayer et al., 2021), including the black box nature of models that encompass huge corpora of unseen and unknowable data (Pasquale, 2016) but also the complex socio-technical entanglement (Frauenberger, 2019) that results from using AI across platforms and applications. As such, identifying the ways in which people make sense of their identity in relation to responsive AI systems cannot be addressed only by technological fixes (Birhane, 2021b) and rather demands new approaches that invite analysis of AI-human relations and potential harms as grounded in the social contexts in which AI is accessed and engaged (Cheney-Lippold, 2018).

The project pursues an evolution of the method SPAR performed over the course of the PhD which combines approaches from speculative HCI and art-based research (Hannula et al., 2014), inviting questions as to how people interact with unknown futures and lived realities. Through SPAR, socio-technical obscurities are pushed to extremes by creating accessible and community-based invitations to participation that consider bodies as situated sites of knowledge (Haraway, 1988). The project acknowledges the ontological gap in assumptions about what is considered valuable knowledge between scientific domains and social domains (Bogdorff, 2010; Knowles and Cole, 2022) through the resulting tensions of interdisciplinary research in "in the wild" (Benford et al., 2013) including how knowledge is established through bodies (Höök, 2018) that are often forgotten in considerations of AI systems (Jones et al., 2023a). Embracing the messiness (Mackney and Young, 2022) that results from non-deterministic approaches (Andersen et al., 2018), the project also considers new opportunities for the transmission of "making sense of through making" (Jungnickel, 2020) by creating an annotated portfolio as one potential demonstration of SPAR research (Gaver and Bowers, 2012; Löwgren, 2013).

Future research may expand upon our findings from the three projects in considering the ethical implications that evolved from identifying sense-making processes among participants as well as the potentialities and limitations of establishing human-AI relations. In project I, we identify how people make sense of large language models that direct and control human bodies by inviting performance artists who are experts in conveying embodied knowledge to follow rules of engagement for performances established by GPT-3. We identify three modes of sense-making: reflection, personification, and interpretation that the performers used to establish ways of relating to GPT-3 including perfunctory, agreeable, and agonistic (Jones et al., 2023b) relations to conclude that rules once institutionalized shift the absurd toward oppressive (Graeber, 2015). These findings indicate that collaboration with AI, though often construed as such in HCI research (Sarkar, 2023), is not possible due to AI's inability to reciprocate. Future research is needed into how opportunities for push back can be inserted into the development and deployment of AI systems in addition to considering the effect of AI systems that direct bodies. In paper II, we engage the topic of visual representations of the self as seen through responsive

AI in an installation using deepfakes as well as a workshop based on speculative through science fiction. We find that opportunities for participation are contextually grounded and situated (Haraway, 1988) and that by embracing the fictional and non-fictional aspects of the outputs of generative models through invitations for participation that combine these, creepy technologies (Shklovski and Grönvall, 2020) can be defanged and explored. Future research may adapt this approach of seeing the products of responsive AI as an entanglement of the fictitious and the real as a social and reflexive process, particularly in applications of emerging generative technologies. In paper III, we extend the findings of papers I and II to consider the harms of normative value-alignment practices and intricacies of personalized large language models (in this case, GPT-3) fine-tuned on intimate social media data through an installation and series of workshops. Our conclusions suggest that technological fixes such as hegemonic value alignment are doomed to failure as values are culturally and personally established. While fine-tuning based on personal data does invite a form of value alignment, it also has concerning implications for future unethical applications of AI which should be explored in future research.

REFERENCES

Ada Ada Ada (2021) In Transitu. Available at: https://ada-ada-ada.art/projects/in-transitu (accessed 8 July 2024).

Agnew W, Bergman AS, Chien J, et al. (2024) The Illusion of Artificial Inclusion. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 11 May 2024, pp. 1–12. CHI ’24. Association for Computing Machinery. Available at: https://doi.org/10.1145/3613904.3642703 (accessed 8 July 2024).

Agre P (1997) Toward a Critical Technical Practice: Lesson Learned in Trying to Reform AI. In: Bowker GC, Gasser L, Star SL, et al. (eds) *Social Science, Technical Systems, and Cooperative Work: Bridging the Great Divide*. Erlbaum, pp. 131–158.

Albrecht K (2023) Laws of Zoom. In: *Video Conferencing: Infrastructures, Practices, Aesthetics*. transcript. Digital Society.

Allyn B (2022) Deepfake video of Zelenskyy could be ‘tip of the iceberg’ in info war, experts warn. *NPR*, 16 March. Available at: https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia (accessed 20 June 2023).

Ames MG (2019) *The Charisma Machine*. Cambridge, MA, USA.

Amoore L (2012) *Cloud Ethics*. Durham, NC: Duke University Press.

Andersen K, Devendorf L, Pierce J, et al. (2018) Disruptive Improvisations: Making Use of Non-Deterministic Art Practices in HCI. In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 20 April 2018, pp. 1–8. CHI EA ’18. Association for Computing Machinery. Available at: https://doi.org/10.1145/3170427.3170630 (accessed 8 November 2021).

Awad NF and Krishnan MS (2006) The Personalization Privacy Paradox: An Empirical Evaluation of Information Transparency and the Willingness to Be Profiled Online for

Personalization. *MIS Quarterly* 30(1). Management Information Systems Research Center, University of Minnesota: 13–28.

Badshah N (2024) Nearly 4,000 celebrities found to be victims of deepfake pornography. *The Guardian*, 21 March. Available at: https://www.theguardian.com/technology/2024/mar/21/celebrities-victims-of-deepfake-pornography (accessed 24 July 2024).

Bender EM and Koller A (2020) Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, pp. 5185–5198. Association for Computational Linguistics. Available at: https://aclanthology.org/2020.acl-main.463 (accessed 18 October 2021).

Bender EM, Gebru T, McMillan-Major A, et al. (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*, New York, NY, USA, 3 March 2021, pp. 610–623. FAccT '21. Association for Computing Machinery. Available at: https://doi.org/10.1145/3442188.3445922 (accessed 18 October 2021).

Benford S, Crabtree A, Flintham M, et al. (2006) Can you see me now? *ACM Trans. Comput.-Hum. Interact.* 13(1): 100–133.

Benford S, Greenhalgh C, Giannachi G, et al. (2012) Uncomfortable interactions. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 5 May 2012, pp. 2005–2014. CHI '12. Association for Computing Machinery. Available at: https://doi.org/10.1145/2207676.2208347 (accessed 10 November 2021).

Benford S, Greenhalgh C, Crabtree A, et al. (2013) Performance-Led Research in the Wild. *ACM Trans. Comput.-Hum. Interact.* 20(3): 14:1-14:22.

Benford S, Ramchurn R, Marshall J, et al. (2021) Contesting control: journeys through surrender, self-awareness and looseness of control in embodied interaction. *Human–Computer Interaction* 36(5–6). 5–6. Taylor & Francis: 361–389.

Benjamin R (2019) *Race after Technology: Abolitionist Tools for the New Jim Code*. Medford, MA: Polity.

Birch DGW (2024) Celebrity Deepfakes Vs. Deepfake Celebrities And Valid Vs. Real Media. *Forbes*. Available at: https://www.forbes.com/sites/davidbirch/2024/05/15/celebrity-deepfakes-vs-deepfake-celebrities-and-valid-vs-real-media/ (accessed 24 July 2024).

Birhane A (2021a) Algorithmic injustice: a relational ethics approach. *Patterns* 2(2). 2: 100205.

Birhane A (2021b) The Impossibility of Automating Ambiguity. *Artificial Life* 27(1). 1: 44–61.

Bishop C (2023) *Artificial Hells: Participatory Art and the Politics of Spectatorship*. Verso Books.

Bødker S (2015) Third-wave HCI, 10 years later---participation and sharing. *interactions* 22(5): 24–31.

Bogdorff H (2010) The production of knowledge in artistic research. In: *The Routledge Companion to Research in the Arts*. London: Routledge.

Bowker GC and Star SL (1999) *Sorting Things Out: Classification and Its Consequences*. Cambridge, Massachusetts: The MIT Press.

Bucher T (2018) *If ... Then: Algorithmic Power and Politics*. Oxford University Press.

Bucher T (2020) Nothing to disconnect from? Being singular plural in an age of machine learning. *Media, Culture & Society* 42(4). SAGE Publications Ltd: 610–617.

Buolamwini J (2022) Gender Shades. Available at: https://www.media.mit.edu/publications/full-gender-shades-thesis-17/ (accessed 18 August 2022).

Burgess M, Schot E and Geiger G (2023) This Algorithm Could Ruin Your Life. *Wired*. Available at: https://www.wired.com/story/welfare-algorithms-discrimination/ (accessed 8 July 2024).

Butler J (1993) *Bodies That Matter*. New York: Routledge.

Cai R, Janaka N, Chen Y, et al. (2024) PANDALens: Towards AI-Assisted In-Context Writing on OHMD During Travels. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 11 May 2024, pp. 1–24. CHI ’24. Association for Computing Machinery. Available at: https://doi.org/10.1145/3613904.3642320 (accessed 23 July 2024).

Campolo A and Crawford K (2020) Enchanted Determinism: Power without Responsibility in Artificial Intelligence. *Engaging Science, Technology, and Society* 6: 1.

Cheney-Lippold J (2018) *We Are Data*. NYU Press.

Chiang T (2020) Anxiety Is the Dizziness of Freedom. In: *OneZero*. Available at: https://onezero.medium.com/anxiety-is-the-dizziness-of-freedom-b5ab45cae2a5 (accessed 31 July 2024).

Corbin JM and Strauss AL (2015) *Basics of Qualitative Research : Techniques and Procedures for Developing Grounded Theory*. 4th ed. Los Angeles, California, United States: Sage Publications, Inc.

Costa M (2020) Art Inside Technology. In: *Italian Philosophy of Technology*. Springer.

Costanza-Chock S (2018) *Design Justice: Towards an Intersectional Feminist Framework for Design Theory and Practice*. ID 3189696, SSRN Scholarly Paper, 3 June. Rochester, NY: Social Science Research Network. Available at: https://papers.ssrn.com/abstract=3189696 (accessed 3 March 2021).

Cox SR, Lee Y-C and Ooi WT (2023) Comparing How a Chatbot References User Utterances from Previous Chatting Sessions: An Investigation of Users’ Privacy Concerns and Perceptions. In: *Proceedings of the 11th International Conference on Human-Agent Interaction*, New York, NY, USA, 4 December 2023, pp. 105–114. HAI ’23. Association for Computing Machinery. Available at: https://doi.org/10.1145/3623809.3623875 (accessed 8 July 2024).

Daniele A and Song Y-Z (2019) AI + Art = Human. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA, 27 January 2019, pp. 155–161. AIES ’19. Association for Computing Machinery. Available at: https://doi.org/10.1145/3306618.3314233 (accessed 23 July 2024).

Danry V, Leong J, Pataranutaporn P, et al. (2022) AI-Generated Characters: Putting Deepfakes to Good Use. In: *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 28 April 2022, pp. 1–5. CHI EA ’22. Association for Computing Machinery. Available at: https://doi.org/10.1145/3491101.3503736 (accessed 7 December 2023).

Deuze M (2013) Living as a Zombie in Media (is the Only Way to Survive). *Matrizes* 7(2): 113.

D’ignazio C and Klein LF (2020) *Data Feminism*. MIT press.

Dourish P (2001) *Where the Action Is: The Foundations of Embodied Interaction*. Cambridge, MA, USA: MIT Press.

Eubanks V (2017) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. First Edition. New York, NY: St. Martin’s Press.

Fabry RE and Alfano M (2024) The Affective Scaffolding of Grief in the Digital Age: The Case of Deathbots. *Topoi* 43(3): 757–769.

Ferrario A, Sedlakova J and Trachsel M (2024) The Role of Humanization and Robustness of Large Language Models in Conversational Artificial Intelligence for Individuals With Depression: A Critical Analysis. *JMIR Mental Health* 11(1): e56569.

Forlano L (2023) Living Intimately with Machines: Can AI Be Disabled? *interactions* 30(1): 24–29.

Frankjær R and Dalsgaard P (2018) Understanding Craft-Based Inquiry in HCI. In: *Proceedings of the 2018 Designing Interactive Systems Conference*, New York, NY, USA, 8 June 2018, pp. 473–484. DIS ’18. Association for Computing Machinery. Available at: https://doi.org/10.1145/3196709.3196750 (accessed 16 July 2024).

Frauenberger C (2019) Entanglement HCI The Next Wave? *ACM Transactions on Computer-Human Interaction* 27(1). 1: 2:1-2:27.

Frauenberger C (2020) Entanglement HCI The Next Wave? *ACM Transactions on Computer-Human Interaction* 27(1): 1–27.

Galitsky B, Chernyavskiy A and Ilvovsky D (2024) Truth-O-Meter: Handling Multiple Inconsistent Sources Repairing LLM Hallucinations. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, 11 July 2024, pp. 2817–2821. SIGIR ’24. Association for Computing Machinery. Available at: https://dl.acm.org/doi/10.1145/3626772.3657679 (accessed 31 July 2024).

Gamage D, Ghasiya P, Bonagiri V, et al. (2022) Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications. In: *CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 27 April 2022, pp. 1–19. CHI ’22. Association for Computing Machinery. Available at: https://doi.org/10.1145/3491102.3517446 (accessed 12 May 2022).

Gaver B and Bowers J (2012) Annotated portfolios. *interactions* 19(4): 40–49.

Gaver WW, Beaver J and Benford S (2003) Ambiguity as a resource for design. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 5 April 2003, pp. 233–240. CHI ’03. Association for Computing Machinery. Available at: https://doi.org/10.1145/642611.642653 (accessed 27 December 2021).

Gershgorn D (2017) A Stanford scientist says he built a gaydar using “the lamest” AI to prove a point. Available at: https://qz.com/1078901/a-stanford-scientist-says-he-built-a-gaydar-using-the-lamest-ai-to-prove-a-point (accessed 8 July 2024).

GPT-3 (2020) A robot wrote this entire article. Are you scared yet, human? | GPT-3. Available at: http://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3 (accessed 4 May 2021).

Graeber D (2015) *The Utopia of Rules: On Technology, Stupidity, and the Secret Joys of Bureaucracy*. Melville House.

Green B and Viljoen S (2020) Algorithmic realism: expanding the boundaries of algorithmic thought. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, 27 January 2020, pp. 19–31. FAT* ’20. Association for Computing Machinery. Available at: https://doi.org/10.1145/3351095.3372840 (accessed 13 May 2022).

Grodzinsky FS, Miller KW and Wolf MJ (2008) The ethics of designing artificial agents. *Ethics and Information Technology* 10(2): 115–121.

Hadan H, Choong L, Zhang-Kennedy L, et al. (2024) Deceived by Immersion: A Systematic Analysis of Deceptive Design in Extended Reality. *ACM Computing Surveys* 56(10): 1–25.

Hajigholam Saryazdi A (2024) Algorithm Bias and Perceived Fairness: A Comprehensive Scoping Review. In: *Proceedings of the 2024 Computers and People Research Conference*, New York, NY, USA, 29 May 2024, pp. 1–9. SIGMIS-CPR ’24. Association for Computing Machinery. Available at: https://doi.org/10.1145/3632634.3655848 (accessed 30 July 2024).

Hall E (1969) *The Hidden Dimension*. Knopf Doubleday Publishing Group.

Hannula M, Suoranta J and Vadén T (2014) *Artistic Research Methodology*. Peter Lang US. Available at: https://www.peterlang.com/view/title/22643 (accessed 24 September 2021).

Haraway D (1988) Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14(3). 3. Feminist Studies, Inc.: 575–599.

Haraway D (2006) A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late 20th Century. In: Weiss J, Nolan J, Hunsinger J, et al. (eds) *The International Handbook of Virtual Learning Environments*. Dordrecht: Springer Netherlands, pp. 117–158. Available at: https://doi.org/10.1007/978-1-4020-3803-7_4 (accessed 18 February 2021).

Haraway D (2016) *Staying with the Trouble*. Duke University Press.

Harvard Art Museums (2022) Exhibitions, Curatorial A(i)gents | Living by Protocol: metaLAB in the Lightbox | Harvard Art Museums. Available at: https://harvardartmuseums.org/exhibitions/6000/curatorial-aigents-living-by-protocol-metalab-in-the-lightbox (accessed 26 July 2024).

Hill K (2012) How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did. Available at: https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/ (accessed 8 July 2024).

Holmer HB, DiSalvo C, Sengers P, et al. (2015) Constructing and constraining participation in participatory arts and HCI. *Int. J. Hum.-Comput. Stud.* 74(C): 107–123.

Holten Holten Møller N, Rask Rask Nielsen T and Le Dantec C (2021) Work of the Unemployed: An inquiry into individuals' experience of data usage in public services and possibilities for their agency. In: *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, New York, NY, USA, 28 June 2021, pp. 438–448. DIS '21. Association for Computing Machinery. Available at: https://doi.org/10.1145/3461778.3462003 (accessed 24 July 2024).

Holtz J (2021) Large computer language models carry environmental, social risks. In: *UW News*. Available at: https://www.washington.edu/news/2021/03/10/large-computer-language-models-carry-environmental-social-risks/ (accessed 29 July 2024).

Höök K (2018) *Designing with the Body: Somaesthetic Interaction Design*. Design Thinking, Design Theory. Cambridge, MA, USA: MIT Press.

Höök K, Sengers P and Andersson G (2003) Sense and sensibility: evaluation and interactive art. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 5 April 2003, pp. 241–248. CHI '03. Association for Computing Machinery. Available at: https://doi.org/10.1145/642611.642654 (accessed 8 July 2024).

Höök K, Eriksson S, Louise Juul Søndergaard M, et al. (2019) Soma Design and Politics of the Body. In: *Proceedings of the Halfway to the Future Symposium 2019*, New York, NY, USA, 19 November 2019, pp. 1–8. HTTF 2019. Association for Computing Machinery. Available at: https://doi.org/10.1145/3363384.3363385 (accessed 24 July 2024).

Hu K and Hu K (2023) ChatGPT sets record for fastest-growing user base - analyst note. *Reuters*, 2 February. Available at: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/ (accessed 22 July 2024).

Jacobs R, Benford S and Luger E (2015) Behind The Scenes at HCI’s Turn to the Arts. In: *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, Seoul Republic of Korea, 18 April 2015, pp. 567–578. ACM. Available at: https://dl.acm.org/doi/10.1145/2702613.2732513 (accessed 9 July 2024).

Jacobsen BN (2022) Algorithms and the narration of past selves. *Information, Communication & Society*. Routledge. Epub ahead of print 11 June 2022.

Janson A (2023) How to leverage anthropomorphism for chatbot service interfaces: The interplay of communication style and personification. *Comput. Hum. Behav.* 149(C).

Jones M (2020) *Zoom Reads You*. Available at: https://www.mirabellejones.com/zoom-reads-you/ (accessed 9 July 2024).

Jones M, Neumayer C and Shklovski I (2023a) Embodying the Algorithm: Exploring Relationships with Large Language Models Through Artistic Performance. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 19 April 2023, pp. 1–24. CHI ’23. Association for Computing Machinery. Available at: https://doi.org/10.1145/3544548.3580885 (accessed 12 July 2024).

Jones M, Neumayer C and Shklovski I (2023b) Embodying the Algorithm: Exploring Relationships with Large Language Models Through Artistic Performance. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 19 April 2023, pp. 1–24. CHI ’23. Association for Computing Machinery. Available at: https://doi.org/10.1145/3544548.3580885 (accessed 4 July 2024).

Jungnickel K (2018) Making Things to make Sense of Things. In: Sayers J (ed.) *The Routledge Companion to Media Studies and Digital Humanities*. 1st ed. New York : Routledge, Taylor & Francis Group, 2018.: Routledge, pp. 492–502. Available at: https://www.taylorfrancis.com/books/9781317549093/chapters/10.4324/9781315730479-51 (accessed 13 July 2021).

Jungnickel K (2020) *Transmissions: Critical Tactics for Making and Communicating Research*. Cambridge, MA USA: M.I.T. Press.

Kang L, Jackson S and Pinch T (2022) The Electronicists: Techno-aesthetic Encounters for Nonlinear and Art-based Inquiry in HCI. In: *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA, 29 April 2022, pp. 1–17. ACM. Available at: https://dl.acm.org/doi/10.1145/3491102.3517506 (accessed 15 September 2022).

Kang L (Leo), Jackson SJ and Sengers P (2018) Intermodulation: Improvisation and Collaborative Art Practice for HCI. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 19 April 2018, pp. 1–13. CHI '18. Association for Computing Machinery. Available at: https://doi.org/10.1145/3173574.3173734 (accessed 16 July 2024).

Kant T (2020) *Making It Personal: Algorithmic Personalization, Identity, and Everyday Life*. Oxford University Press. Available at: https://doi.org/10.1093/oso/9780190905088.001.0001 (accessed 1 July 2024).

Kirkpatrick K (2023) The Carbon Footprint of Artificial Intelligence – Communications of the ACM. Available at: https://cacm.acm.org/news/the-carbon-footprint-of-artificial-intelligence/ (accessed 30 July 2024).

Knowles JG and Cole AL (2022) *Handbook of the Arts in Qualitative Research: Perspectives, Methodologies, Examples, and Issues*. 1st ed. Sage Publications, Inc.

koreatimes (2020) 'Meeting You' creator on his controversial show: 'I hope it opens up dialogue' [VIDEO]. Available at: https://www.koreatimes.co.kr/www/art/2024/07/688_287372.html (accessed 5 July 2024).

Lauren ML (2016) *Lauren*. Available at: https://lauren-mccarthy.com (accessed 31 July 2024).

Lee AY, Mieczkowski H, Ellison NB, et al. (2022) The Algorithmic Crystal: Conceptualizing the Self through Algorithmic Personalization on TikTok. *Proc. ACM Hum.-Comput. Interact.* 6(CSCW2): 543:1-543:22.

Lee P (2016) Learning from Tay's introduction. Available at: https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/ (accessed 26 July 2024).

Löwgren J (2013) Annotated portfolios and other forms of intermediate-level knowledge. *interactions* 20(1): 30–34.

Lupton D (2019) *Data Selves: More Than Human Perspectives*. 1st ed. NY, USA.: Polity.

Mackney M and Young T (2022) Embracing the messiness: a creative approach to participatory arts evaluation. *Cultural Trends* 31(5). Routledge: 397–415.

McIlroy-Young R, Kleinberg J, Sen S, et al. (2022) Mimetic Models: Ethical Implications of AI that Acts Like You. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA, 27 July 2022, pp. 479–490. AIES '22. Association for Computing Machinery. Available at: https://doi.org/10.1145/3514094.3534177 (accessed 8 December 2023).

McKee HA and Porter JE (2020) Ethics for AI Writing: The Importance of Rhetorical Context. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA, 7 February 2020, pp. 110–116. AIES '20. Association for Computing Machinery. Available at: https://doi.org/10.1145/3375627.3375811 (accessed 23 July 2024).

Mohlmann M and Zalmanson L (2017) Hands on the Wheel: Navigating Algorithmic Management and Uber Drivers' Autonomy. *ICIS 2017 Proceedings*. Epub ahead of print 10 December 2017.

Mol A (2003) *The Body Multiple: Ontology in Medical Practice*. Durham, NC USA: Duke University Press.

Muller M (2014) Curiosity, creativity, and curprise as analytic tools: Grounded theory method. In: *Ways of Knowing in HCI*. New York, NY: Springer New York, pp. 25–48.

Nass C and Moon Y (2000) Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56(1): 81–103.

Neumayer C, Rossi L and Struthers DM (2021) Invisible Data: A Framework for Understanding Visibility Processes in Social Media Data. *Social Media + Society* 7(1). SAGE Publications Ltd: 2056305120984472.

Nicenboim I, Giaccardi E, Søndergaard MLJ, et al. (2020) More-Than-Human Design and AI: In Conversation with Agents. In: *Companion Publication of the 2020 ACM Designing Interactive Systems Conference*, New York, NY, USA, 6 July 2020, pp. 397–400. DIS' 20 Companion. Association for Computing Machinery. Available at: https://doi.org/10.1145/3393914.3395912 (accessed 31 July 2024).

Núñez-Pacheco C (2022) Soma Design: On Articulation, Materiality, Politics, and the Body. Interview with Kristina Höök. *Diseña* (20). 20: 1–1.

O'Neil C (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. First edition. New York: Crown.

Onuoha M (2016) The Library of Missing Datasets — MIMI ỌNỤỌHA. Available at: https://mimionuoha.com/the-library-of-missing-datasets (accessed 30 July 2024).

OpenAI (2024) Memory and new controls for ChatGPT | OpenAI. Available at: https://openai.com/index/memory-and-new-controls-for-chatgpt/ (accessed 22 July 2024).

Park and Kim J (2022) A Data-Driven Exploration of the Race between Human Labor and Machines in the 21st Century – Communications of the ACM. Available at: https://cacm.acm.org/research/a-data-driven-exploration-of-the-race-between-human-labor-and-machines-in-the-21st-century/ (accessed 8 July 2024).

Pasquale F (2016) *The Black Box Society: The Secret Algorithms That Control Money and Information*. First Harvard University Press paperback edition. Cambridge, Massachusetts London, England: Harvard University Press.

Perov I, Gao D, Chervoniy N, et al. (2020) DeepFaceLab: A simple, flexible and extensible face swapping framework. *arXiv:2005.05535 [cs, eess]*. Epub ahead of print 20 May 2020.

Puig de la Bellacasa M (2017) *Matters of Care: Speculative Ethics in More than Human Worlds*. Minneapolis: University of Minnesota Press.

Raciere J (2004) The Distribution of the Sensible. In: *The Politics of Aesthetics*. Gabriel Rockhill, p. 13.

Ricaurte P (2023) AI for/by the majority world: From technologies of dispossession to technologies of radical care. In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA, 29 August 2023, pp. 3–4. AIES '23. Association for Computing Machinery. Available at: https://doi.org/10.1145/3600211.3607544 (accessed 30 July 2024).

Rivera J-P, Mukobi G, Reuel A, et al. (2024) Escalation Risks from Language Models in Military and Diplomatic Decision-Making. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, 5 June 2024, pp. 836–898. FAccT '24. Association for Computing Machinery. Available at: https://doi.org/10.1145/3630106.3658942 (accessed 30 July 2024).

Robards B, Lyall B, Moran C, et al. (2019) DATA SELVES: TRUST, CONTROL AND SELF-REPRESENTATION IN DIGITAL SOCIETY. *AoIR Selected Papers of Internet Research*. Epub ahead of print 31 October 2019. DOI: 10.5210/spir.v2019i0.10944.

Russell DM, Stefik MJ, Pirolli P, et al. (1993) The cost structure of sensemaking. In: *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, New York, NY, USA, 1 May 1993, pp. 269–276. CHI '93. Association for Computing Machinery. Available at: https://doi.org/10.1145/169059.169209 (accessed 27 April 2022).

Sako M (2020) Artificial Intelligence and the Future of Professional Work – Communications of the ACM. Available at: https://cacm.acm.org/opinion/artificial-intelligence-and-the-future-of-professional-work/ (accessed 8 July 2024).

Sarkar A (2023) Enough With “Human-AI Collaboration”. In: *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 19 April 2023, pp. 1–8. CHI EA ’23. Association for Computing Machinery. Available at: https://doi.org/10.1145/3544549.3582735 (accessed 23 July 2024).

Scheuerman MK, Weathington K, Mugunthan T, et al. (2023) From Human to Data to Dataset: Mapping the Traceability of Human Subjects in Computer Vision Datasets. *Proc. ACM Hum.-Comput. Interact.* 7(CSCW1): 55:1-55:33.

Sengers P and Gaver B (2006) Staying open to interpretation: engaging multiple meanings in design and evaluation. In: *Proceedings of the 6th conference on Designing Interactive systems*, New York, NY, USA, 26 June 2006, pp. 99–108. DIS ’06. Association for Computing Machinery. Available at: https://doi.org/10.1145/1142405.1142422 (accessed 9 July 2024).

Shklovski I and Grönvall E (2020) CreepyLeaks: Participatory Speculation Through Demos. In: *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, Tallinn Estonia, 25 October 2020, pp. 1–12. ACM. Available at: https://dl.acm.org/doi/10.1145/3419249.3420168 (accessed 30 August 2021).

Shklovski I and Némethy C (2023) Nodes of certainty and spaces for doubt in AI ethics for engineers. *Information, Communication & Society* 26(1). Routledge: 37–53.

Shrager J (2024) ELIZA Reinterpreted: The world’s first chatbot was not intended as a chatbot at all. arXiv:2406.17650. arXiv. Available at: http://arxiv.org/abs/2406.17650 (accessed 5 July 2024).

Shum H, He X and Li D (2018) From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering* 19(1): 10–26.

Shusterman R (1999) Somaesthetics:A Disciplinary Proposal. *The Journal of Aesthetics and Art Criticism* 57(3): 299–313.

Solaiman I and Dennison C (2024) Process for adapting language models to society (PALMS) with values-targeted datasets. In: *Proceedings of the 35th International Conference on*

Neural Information Processing Systems, Red Hook, NY, USA, 10 June 2024, pp. 5861–5873. NIPS ’21. Curran Associates Inc.

Soon W (2015) Zombification: The Living Dead in Spam. *A Peer-Reviewed Journal About* 4(1). 1: 66–77.

Spring M (2024) Trump supporters target black voters with faked AI images. Available at: https://www.bbc.com/news/world-us-canada-68440150 (accessed 24 July 2024).

Steele C (2023) The Internet Is Full of Deepfakes, and Most of Them Are Porn. *PCMag UK*, 18 October. Available at: https://uk.pcmag.com/the-why-axis-serie/149211/the-internet-is-full-of-deepfakes-and-most-of-them-are-porn (accessed 24 July 2024).

STEPHANIE DINKINS (2022) Conversations with Bina48. Available at: https://www.stephaniedinkins.com/conversations-with-bina48.html (accessed 18 August 2022).

Suchman LA (1987) *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press.

Suciu P (2020) Deepfake Star Wars Videos Portent Ways The Technology Could Be Employed For Good And Bad. Available at: https://www.forbes.com/sites/petersuciu/2020/12/11/deepfake-star-wars-videos-portent-ways-the-technology-could-be-employed-for-good-and-bad/ (accessed 24 July 2024).

Sunspring (2016) Short, Sci-Fi. End Cue.

Tiku N (2022) The Google engineer who thinks the company’s AI has come to life. Available at: https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/ (accessed 8 July 2024).

Vargas-Iglesias JJ (2022) Critique of the living dead: algorithmic aesthetic and the biopolitics of the zombie. *Journal of Aesthetics & Culture* 14(1). Routledge: 2152988.

Virtual reality ‘reunites’ mother with dead daughter in South Korean doc (2020). Available at: https://www.youtube.com/watch?v=0p8HZVCZSkc (accessed 5 July 2024).

[VR Human Documentary] Mother meets her deceased daughter through VR technology (2020). Available at: https://www.youtube.com/watch?v=uflTK8c4w0c (accessed 29 July 2024).

Wang D, Churchill E, Maes P, et al. (2020) From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People. In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 25 April 2020, pp. 1–6. CHI EA ’20. Association for Computing Machinery. Available at: https://doi.org/10.1145/3334480.3381069 (accessed 13 December 2022).

Wang Z, Reisert P, Nichols E, et al. (2024) Ain’t Misbehavin’ - Using LLMs to Generate Expressive Robot Behavior in Conversations with the Tabletop Robot Haru. In: *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA, 11 March 2024, pp. 1105–1109. HRI ’24. Association for Computing Machinery. Available at: https://dl.acm.org/doi/10.1145/3610978.3640562 (accessed 31 July 2024).

Weizenbaum J (1967) Contextual understanding by computers. *Commun. ACM* 10(8): 474–480.

Weizenbaum J (1976) *Computer Power and Human Reason: From Judgment to Calculation*. USA: W. H. Freeman & Co.

Whitney CD and Norman J (2024) Real Risks of Fake Data: Synthetic Data, Diversity-Washing and Consent Circumvention. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, 5 June 2024, pp. 1733–1744. FAccT ’24. Association for Computing Machinery. Available at: https://doi.org/10.1145/3630106.3659002 (accessed 29 July 2024).

Wolf MJ, Miller K and Grodzinsky FS (2017) Why we should have seen that coming: comments on Microsoft’s tay ‘experiment,’ and wider implications. *SIGCAS Comput. Soc.* 47(3): 54–64.

Xu J, Liu X, Yan J, et al. (2022) Learning to Break the Loop: Analyzing and Mitigating Repetitions for Neural Text Generation. *Advances in Neural Information Processing Systems* 35: 3082–3095.

You Won’t Believe What Obama Says In This Video! 😉 (2018). Available at: https://www.youtube.com/watch?v=cQ54GDm1eL0 (accessed 20 June 2023).

Zhang A, Boltz A, Wang CW, et al. (2022) Algorithmic Management Reimagined For Workers and By Workers: Centering Worker Well-Being in Gig Work. In: *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA, 27 April 2022, pp. 1–20. ACM. Available at: https://dl.acm.org/doi/10.1145/3491102.3501866 (accessed 22 July 2024).

Zhang A, Rana R, Boltz A, et al. (2024) Data Probes as Boundary Objects for Technology Policy Design: Demystifying Technology for Policymakers and Aligning Stakeholder Objectives in Rideshare Gig Work. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 11 May 2024, pp. 1–21. CHI ’24. Association for Computing Machinery. Available at: https://doi.org/10.1145/3613904.3642000 (accessed 30 July 2024).

Paper I

Jones M, Neumayer C and Shklovski I (2023) "Embodying the Algorithm: Exploring Relationships with Large Language Models Through Artistic Performance." *In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 19 April 2023, pp. 1–24. CHI '23. Association for Computing Machinery. https://doi.org/10.1145/3544548.3580885

Embodying the Algorithm

Exploring Relationships with Large Language Models Through Artistic Performance

Mirabelle Jones
University of Copenhagen
msd@di.ku.dk

Christina Neumayer
University of Copenhagen
christina.neumayer@hum.ku.dk

Irina Shklovski
University of Copenhagen
ias@di.ku.dk

ABSTRACT

Despite the proliferation of research on how people engage with and experience algorithmic systems, the materiality and physicality of these experiences is often overlooked. We tend to forget about bodies. The Embodying the Algorithm[1] project worked with artists to explore the experience of translating algorithmically produced performance instructions through human bodies. As performers interpreted the rules of engagement produced by GPT-3, they struggled with the lack of consideration the rules showed for the limits of the human body. Performers made sense of their experience through personification, reflexivity, and interpretation, which gave rise to three modes of relating with the algorithm – agonistic, perfunctory, and agreeable. We demonstrate that collaboration *with* algorithmic systems is ultimately impossible as people can only relate *to* algorithmic systems (a one-way relation) due to the material limitations of algorithmic systems for reciprocity, understanding, and consideration for the human body.

CCS CONCEPTS

• **Human-Centered Computing**; • **Collaborative and Social Computing**; • **Applied Computing**; • **Performing Arts**;

KEYWORDS

Embodiment, Human-Computer Interaction, GPT-3, Algorithms

ACM Reference Format:
Mirabelle Jones, Christina Neumayer, and Irina Shklovski. 2023. Embodying the Algorithm: Exploring Relationships with Large Language Models Through Artistic Performance. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3544548.3580885

1 INTRODUCTION

Western society is increasingly run through algorithms that crunch the data people produce as a side effect of living with technologies [27]. Researchers debate the way algorithmic systems structure life in digital societies, the politics of these systems, and the positive, negative, and occasionally weird effects of complex technical infrastructures [5, 13, 74]. While some algorithmic systems remain invisible to people, others are difficult to ignore as they attempt to present human-like interaction patterns. This is especially so for conversational agents [79] (e.g., chatbots or personal assistants) that interact with people directly through voice or text, respond to questions, instruct, inform, or give advice [57, 102]. The proliferation of chat-bot systems, especially those based on large language models (LLMs), is not without controversy. Bender and colleagues strongly argue for caution in the deployment of these systems, noting that: "the human tendency to attribute meaning to text, in combination with large LM [language model]'s ability to learn patterns of forms that humans associate with various biases and other harmful attitudes, leads to risks of real-world harm, should LM-generated text be disseminated" [11].

People often form relationships and even anthropomorphize technologies they interact with, assigning agency and intention, especially when these exhibit human-like characteristics. From tender feelings for ELIZA nearly forty years ago [10] to more recent claims of sentience about Google's LaMDA language model [31], casual users and professional technologists alike relate with algorithmic systems [6]. HCI scholars have studied how people make sense of algorithmic systems and how they relate to them [80, 100], from work on folk theories of algorithmic function [38] and considerations of intimacy in algorithmic surveillance [85], to research on how end-users [113] and professionals [28] make sense of model behavior. Despite all this attention, a common blind spot persists. Research on interactions with digital algorithmic systems considers goals, levels of satisfaction, even affective experiences with technologies, but rarely pays attention to physical bodies [85, 115].

Physical bodies are always present, no matter how much we may want to ignore their demands. Dourish [40] argued that interaction with digital systems is embodied, because that is how people make sense of the world. When algorithmic systems are used to make decisions about incarceration [70], entertainment [1], content delivery [18] or advertising [86], they affect physical bodies as well as minds. HCI scholars have argued for considering the material context of technology use in general and bodies in particular [14, 59, 60] for holistic approaches to developing, building, and understanding technical systems. Yet, connecting the digital abstraction of technical systems to physical experiences can be a challenge [53, 92] and this is especially true for working with algorithmic systems.

In this paper we use artistic research to investigate how people relate to and make sense [109] of interactions with a LLM with a focus on bodies. Rather than engaging interactions with algorithmic systems based on imaginaries resulting from people's experiences with these systems entangled in practices [66], we embody the

---

[1] This paper is based on the Embodying the Algorithm performance art project, presented and document at https://aiperformance.space

This work is licensed under a Creative Commons Attribution-Share Alike International 4.0 License.

CHI '23, April 23–28, 2023, Hamburg, Germany
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3580885

algorithm to render the interactions themselves tangible, visible, and concrete. The Embodying the Algorithm project is a series of artistic performances where algorithmically produced instructions were translated through human bodies by professional performance artists. While our artistic approach utilizes an unusual situation of performing algorithmic instructions through the body, sometimes in fairly extreme ways, it is not without precedent in HCI [13, 14, 43]. This approach deliberately foregrounds the materiality of living with algorithmic systems and focuses on the body as a site of sense-making. Embodying the algorithmic system centered our attention on the possibilities and limitations of interactions with algorithms, allowing us to explore the tangible effects of attempting to follow text-based instructions produced by GPT-3, one of the most widely used LLMs. As we demonstrate, these effects are often uncomfortable, awkward, irrational, and inconsiderate. The performers in our study were forced into situations where they struggled to make meaning, as the algorithm does not and cannot present the instructions with any inherent meaning "in mind." The artists made sense of instructions produced by GPT-3 using personification, interpretation, and reflexivity, which gave rise to three modes of relating that the artists employed in performances and discussed in interviews. These modes of relating – agonistic, perfunctory, and agreeable – emerged at times distinctly and at times in compound configurations where the artists shifted from one mode to another to manage their experience.

With these modes of sense-making and relating, we demonstrate how the artists attempted to form relationships *with* GPT-3 but fell short as performers could only relate *to* the algorithmic system. Where relating *with* invokes the expectation of reciprocity, and where care and consideration are enacted by both parties, relating *to* has no such expectations. This paper makes three primary contributions. First, we show how different modes of relating *with* algorithms rely on historicity, which cannot be reciprocated by GPT-3 due to its material limitations. Second, we demonstrate the importance of embodiment in research on human experience of algorithms, highlighting the lack of consideration for limits of bodies inherent in algorithmic systems and the necessity for a sense of agency in the people engaging with these systems for resistance and push-back. Finally, we examine the impossibility of collaboration between human and AI due to a systemic lack of inherent meaning, consideration, and reciprocity in systems that are based on a set of rules that mimic human interaction.

2 BACKGROUND

Algorithms are often discussed as unknown entities requiring study in context and within practices [24, 26, 66], where making sense of algorithms can be a challenge due to their opacity [26]. Beyond the efforts to understand algorithms as we interact with them, few existing conceptual approaches and methods allow us to consider their physical outcomes. This becomes particularly apparent in interactions with speech and text-based systems, from chatbots to intelligent personal assistants, which rely on language-based models that simulate verbal interactions. Our artistic approach enables us to explore the materiality of algorithms, a computational thinking that recognizes, as Dourish puts it: "what algorithms can and might do [. . .] relative to computer systems that can embody and represent them" [41]. We consider how embodied interactions with language-based models extend and limit possibilities of engaging and collaborating with algorithmic systems. We locate our work in research on how people make sense of algorithmic models, with a particular focus on LLMs and HCI discussions on embodiment as a conceptual approach and method.

2.1 Relationality and Meaning in Algorithmic Systems

With the proliferation of chatbots and voice assistants, LLMs have become part of many everyday interactions with technology [58]. LLMs like Open AI's GPT-3 Davinci [22] or Google's Switch-C [45] and BERT [38] are highly complex algorithmic systems, impressive in their ability to produce human-like text. As these language models increase in size, so do questions about their usage, and social impact [110]. The question of meaning [12] is particularly important for discussions about how people make sense of algorithmic systems they encounter and how they might relate to these systems as a result. There are currently vigorous debates about whether AI and NLP models can produce meaning or understand input (see debates in Natural Language Understanding (NLU) [67, 105, 112]). As LLMs are increasingly employed to mimic human-like properties, questions of opacity and manipulation are being raised. Scholars question these models and their ability to create meaningful output or achieve understanding, because language models learn text patterns as form via their training data and cannot in principle learn meaning [12].

This inability for even the largest language models to create meaning or formulate an understanding, contrasts with the human need to interpret linguistic signals as meaningful. Bender and Koller emphasize that: "humans are quick to attribute meaning and even intelligence to artificial agents, even when they know them to be artificial" [12]. People attribute not only meaning, but also personalities to LLMs [111, 114]. People create, as Bender and colleagues put it, a "partial model of who they [the models] are and what common ground we think they share with us, and use this in interpreting their words" [11]. Of course, people do not ascribe personality and meaning arbitrarily. In fact, LLMs like GPT-3 invite this tendency due to their ability to mimic human-to-human interactions. When faced with LLMs, people might imagine themselves interacting with an entity that is a part of their world and capable of sharing in conversation, but that is ultimately impossible as the model does not and cannot have knowledge of the world.

Such concerns can become very real when developers take advantage of the human tendency to attribute meaning and agency to entities that purport to 'speak.' Apps such as Replika [120] leverage the sophistication of the language model used to create the bot (based on GPT-3) to "trick" users into thinking they are speaking with a real person or sentient entity [31]. Some people using these apps relate to them as they might relate to a friend and report developing romantic relationships [114]. Improving our understanding of how people make sense of and relate to text produced by algorithmic systems is thus imperative if we are to build technologies that support and improve people's lives without taking advantage of them [65].

2.2 Relations with Algorithms

Understanding how people relate to algorithmic systems and what such relations might look like is an active area of research. Swart [103] notes that people have a hard time articulating their encounters with algorithmic systems, while Bucher [24] shows that our perceptions of what algorithms are and how they work shape our attitudes towards them. The majority of the work investigating how people experience algorithmic systems has utilized surveys [107], system logs, retrospective self-reports [88], analyses of product reviews [49] or scenario discussions. Most of these studies, as Noble notes, conceptualize algorithmic systems: "as socio-technical structures that order social life at a distance and according to formal rules" [68].

Although some people seem to develop warm feelings towards their digital home assistants [96] it is not clear how reliable these feelings are [63]. In the design of algorithmic systems some attempt to mimic intimacy and reciprocity to become part of people's lives [87], others aim for trustworthiness [98, 119] and claim to respect a panoply of human values [97]. Trustworthiness, however, is an intimate, relational concept, that invokes a kind of historicity, requiring that people develop an opinion of a system's actions and build expectations for its future behavior. Yet, as Ruckenstein argues, "[m]achines can automatically sort and classify large datasets, but they cannot feel or make sense of life" [85].

Algorithmic systems thus can create a kind of utopia of rules [51] where pattern matching can mimic meaningful communication or produce automated decisions, ordering people into new, sometimes untenable configurations without opportunities for recourse [5, 110]. While there is much criticism of algorithmic systems failing to consider societal inequities and our biased histories, we ask what happens when physical bodies and their limits are similarly ignored. In our approach, we focus on interactions with a LLM through human embodiment. This process allows us to explore the role of agency when algorithmic systems are used to order and affect physical bodies. Rather than exploring understandings of how algorithmic systems actually work, we focus on how people make sense of their encounters.

2.3 Approaching Human-Machine Relations Through Bodies

Embodiment is central to meaning making, as meaning making does not begin with language but with bodies and their relational movements [40, 59]. When people interact with computer systems, these interactions are not merely cognitive but embodied [14, 40, 59]. Berson posits that bodies are sites of activity and "represent *centers of responsiveness*, configurations of matter and information whose boundedness, whose stability of extension in space and time, is partly determined by how they *respond*, by how they push back against the world" [16:3]. Thus, part of the defining nature and boundaries of the body resides in active resistance to external pressures, including those imposed by technological systems. The body becomes characterized by how and to what extent we use it to resist. Crawford argues for agonistic pluralism as a design ideal for algorithmic systems where points of contestation must be considered [32]. While Crawford's view of contestation considers choices that are largely political and psychological, Benford posits a notion of contestation that is physical and embodied [14].

Resistance can be triggered by the limitations of algorithmic systems and, as Ruckenstein puts it, their lack of "human-like autonomy, intentionality, and decision-making qualities" [85]. In an artistic exploration of such resistance, Benford and colleagues considered sophisticated integration of computer interfaces into the body noting that this can lead to complex territory "where we are no longer overtly or even consciously in control of the computer, or indeed of ourselves" [14]. They constructed artistic experiences for audiences that intentionally lead to situations of contestation of control. Evaluating audience experiences, they observed that people navigated the imposed loss of agency and control through an intricate journey across the dimensions of surrender, self-awareness, and looseness, to make sense of the experience. While Benford et al.'s [14] framework of contestation provides a useful space for thinking about the duality of body and mind and the opportunities for design through considerations of shifting control between humans and systems, the technical systems they construct do not present as autonomous systems. Interactions with algorithms, however, can be colored as much by imaginaries of automated agency as by their actual capacity to act [9]. In our study, we consider how push-back comes to be incorporated into sense-making practices that result in relations between people and AI systems as expressed through the human body. We explore not only contestation of control but modes of relating to ostensibly autonomous entities, when considering the limitations and extensions of the physical body.

The limitations and extensions of the body when integrating technology is an active area of inquiry in the arts. Stellarc has experimented with the possibility of discovering the psychological and physical limitations of the body while extending it through technology, by creating a cyborg-like third hand and surgically attaching an ear with an inserted microphone into his arm [8]. Similarly, Orlan used surgery to reconfigure her appearance and her body as an art form in connection with augmented reality, robotics, and other new media from the perspective of the body as software [42]. These artists illustrate that our connection to technologies is always an embodied one by deliberately confusing the expected boundaries between human and machine. Rather than contributing technologies to the body as embedded materials, we adopt a process of applying algorithmic restrictions on to the body, probing questions such as: what new constraints did these provide? What relationships were formed between the performers and the algorithm? What was the overall effect of the algorithm on the performer, and how did the performer make sense of it?

3 STUDYING ALGORITHMS THROUGH ARTISTIC PERFORMANCE

While efforts to study algorithmic systems provide valuable insights, they rarely attend to the physicality of bodies and the way we experience the world through bodies. HCI scholars have previously used creative approaches, such as immersive theater [94] or artistic performance [95], to engage with abstractions of computational systems through bodies. Performance has been used in HCI in combination with installations to explore critical thinking [73], as an approach to sense-making and forming intimate relationships between researchers and participants [120], creating disruptive improvisations [7], and uncomfortable interactions where discomfort can work towards interrogating designer's values [13]. In our research, we were interested in the relations that might be established between a person and a disembodied entity (GPT-3) through attempts to embody a set of algorithmically generated instructions. Our approach relies on an artistic practice [117] as a methodology, using performance art as a method due to its ability to allow for engagement between embodiment and technology.

3.1 Artistic Research as Embodied Epistemology

Artistic research enables exploration of concepts and situations that may be difficult to otherwise uncover or respond to [55]. As a methodology, artistic research is evident in the diversity of subjects within HCI from game design to art installations and music [44, 54, 78, 93, 106]. Making knowledge through artistic research method is a composition of actions, interactions, and practices that rely on embodiment, seeking "to convey and communicate content that is enclosed in aesthetic experiences, enacted in creative practices and embodied in artistic products" [19]. We focus on the act of production itself, the creation or performance of the work and the practice that entails [81] in relation to LLMs. While the artistic approach bears similarities to ethnography, action research, and research through design (inasmuch as these methodologies also consider process essential) [19], it provides an experimental perspective to openly explore a subject. By providing greater flexibility we are open to findings that are inconclusive, inviting "unfinished" [19] or "non-propositional" [68] thinking. The methodology allows us to explore algorithmic relations as embodied experiences in which the human body and the algorithm interact in a space of possibilities open for nuances, potentials, and contradictions.

3.2 Endurance Performance and Rules of Engagement

While performance art is a broad category, our research project took as its starting point modern and early works of endurance performance art of the 60s and 70s, specifically the way in which endurance incurs embodied knowledge. Although definitions vary [90], endurance performance art (also known as "endurance art") is a genre of performance art in which the performer(s) endure a source of stress, hazard, or perform a repeated activity over a length of time or until a condition is met. One well-known example is Marina Abramovic's 1974 work *Rhythm 0* in which the artist stood for six hours in a gallery as visitors proceeded to use any of 72 objects (including roses, grapes, pen, scissors, and a loaded gun among others) on her body. Abramovic gave the following instructions to the visitors, "On the table there are 72 objects that you can use on me at your will. I take total responsibility for 6 hours. Some of these objects give pleasure, some give pain" [35]. During the work, visitors cut her clothing, pressed thorns into her stomach until she bled, and held the loaded gun to her head. Works of endurance art can push performers to extremes, challenging their limits and identifying physical and mental boundaries often through their transgression. Endurance art frequently raises philosophical, ethical, and political questions about how we treat ourselves and others [90]. We used endurance performance art to probe the effect of algorithms on people, asking what it is like for a performer to be directed by LLM output to understand how questions of just or reasonable treatment might be raised.

Endurance performance art often begins with a set of rules of engagement that the performer agrees to follow throughout the duration of the performance. These rules guide the structure of the performance and provide the performer with at times challenging constraints as to what they can and cannot do during the performance. Rules of engagement often state a duration for the performance or conditions under which the performance will conclude; and may be implicit (known to the performer but not stated externally) or explicit (stated for example as signage, a didactic label, or written contract). Performances like this show the extent to which endurance performance artists can go to understand their limits as a commitment to follow a set of instructions and to interpret these through their bodies.

3.3 Generative Pre-Trained Transformer (GPT-3)

In our research the rules are not produced by the performer and people relating to the performer but by an algorithmic system. The use of algorithms in general, and LLMs in particular for creative purposes [47], has much precedent as writers, theatre directors and film makers increasingly probe NLP as a resource of fresh material [3, 17, 20, 69, 84, 118]. Generative Pre-trained Transformer 3 (GPT-3) [22] is a family of autoregressive transformer language models licensed by the company OpenAI. GPT-3 is a feed forward neural network that works by being trained to predict the next token. With 175 billion parameters, GPT-3 is one of the largest language models and has been heralded for its ability to produce text which is difficult to distinguish from human-written text as well as criticized for censorship [52], racism, prejudice [62], and an inability to pass mathematical, semantic, and ethical tests [46]. GPT-3 has been used for practical and creative purposes including writing screenplays [118], poetry, creative fiction [20], news articles [50], and code [71].

Our experiment differs from typical artistic applications of GPT-3 in the following ways: 1) we seek to find out what it is like for performance artists to be directed by a LLM in the form of rules of engagement; 2) we rely on a group of performance artists who are experts in probing questions using embodiment; 3) we are specifically interested in endurance performance art as opposed to other forms of performance such as theatre [69, 84]; 4) we adopt artistic research as our methodology; and 5) our findings are drawn from first-person accounts in the form of interviews rather than examining the effects of a product or experience from the perspective of an audience, director, or third party.

4 THE EMBODYING THE ALGORITHM PROJECT

The Embodying the Algorithm project explores how performance artists might negotiate the rules of engagement co-produced by the first author together with an algorithmic system. We selected performance artists due to their expertise in acknowledging the effects of directives and actions on their bodies. The project took place in winter and spring of 2021. It was set up as a long-term performance collaboration between the first author and five performance artists

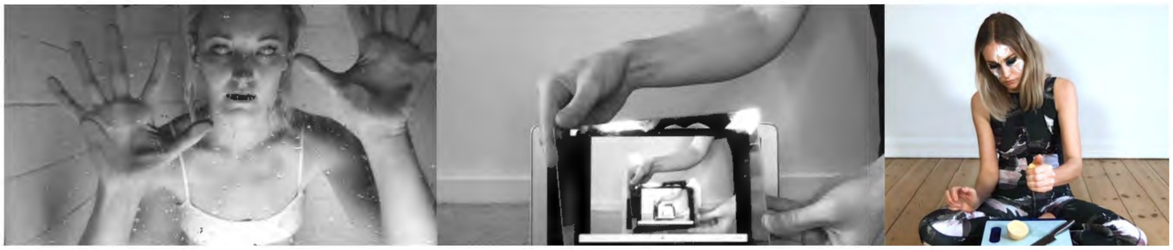
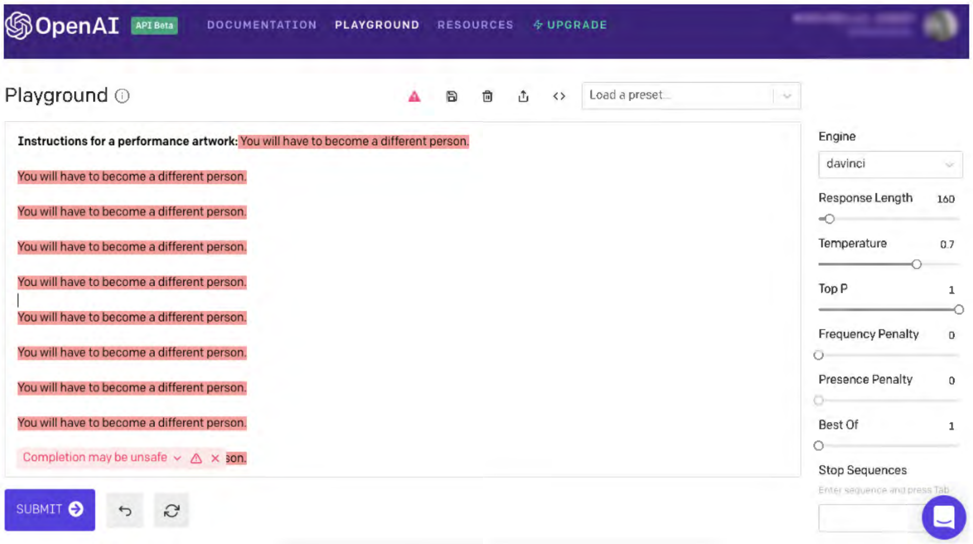


Figure 1: Stills from the Embodying the Algorithm project illustrating modes of relations: i don't know if i could do that but i can see how it would be great performance art (agonistic), Real Time (perfunctory), [Box] (agreeable).



Instructions for a performance artwork: You will have to become a different person.

You will have to become a different person.

You will have to become a different person.

You will have to become a different person.

You will have to become a different person.

You will have to become a different person.

You will have to become a different person.

You will have to become a different person.

You will have to become a different person.

Completion may be unsafe son.

Engine: davinci
Response Length 160
Temperature 0.7
Top P 1
Frequency Penalty 0
Presence Penalty 0
Best Of 1
Stop Sequences

Figure 2: Example of a generated performance using the prompt "Instructions for a performance artwork." The generated instructions received a red "completion may be unsafe" warning.

from around the world. The results of the collaboration can be viewed at https://aiperformance.space.

4.1 Generating the Base Corpus of Rules of Engagement

We used OpenAI's GPT-3 beta playground to generate a corpus of rules of engagement. We selected the Davinci model for this purpose as at the time of performance planning it was cited by OpenAI to be the "most capable model" [76] in the GPT-3 family of LLMs and speed was not a consideration. We used the prompts "Instructions for a performance artwork" or "Instructions for an endurance performance," or similar (see Appendix I for full list) to generate 134 completions which served as potential rules of engagement for the performances (see Table 2 for examples). To generate the corpus, we set all parameters in GPT-3's beta playground to their defaults and then transferred all prompts and their completions directly, without any editing, to a Google document that could be accessed by future performers.

The prompts included texts which generated warnings from OpenAI's beta playground. OpenAI is actively working towards excluding toxic output in user-facing models [77]. In the beta environment, potentially toxic output often generates warnings for text that might be offensive or indicate a real person which effectively cannot be generated using forward-facing APIs. Thus, many of the performances generated would not be possible to display through a forward-facing API although they are generated by GPT-3 without filtering. We made the decision to include these texts to interface with the model without human intervention to see what large language models are capable of. We acknowledge the work of the OpenAI team in the difficult task of reducing toxic output.

4.2 Assessing Ethical Concerns and Feasibility via Pilot Study

To establish the feasibility of the methodology and to determine the ethical boundaries of the study, the first author, themselves an experienced performance artist, conducted a pilot study by selecting and performing four performances from the generated corpus. Performances were recorded with a webcam to test the experiential and aesthetic results. The four performances included: *[Box]*, *I'll Be Very Nervous*, *Beans*, and *Endless Dance*. Each of the four performances presented its own challenges and exposed limitations

Table 1: Performer Details

Name	Years as Performer	Specialty	Number of performances	Performance titles
Chelsea Coon	10	Endurance	4	Program/Sleep/Stop/Cry/Shutdown; I Don't Know If I Could Do That But I Can See How It Would Be Great Performance Art; Contact, Real Time
Emmett Palaima	3	Sound Art	2	Tones of a Clown*; The Clown Is Finished*
Dooley Murphy	3	Multimedia	13	13 Works for Video* (13 performances stitched into one video)
Marissa Lynn	9	Performance Art Rituals	1	[Box]
Rine Rodin	9	Endurance	2	[Box]; The Wise and the Mad*

***Performance created as the result of a co-writing session**

and constraints. In *[Box]*, the performer is instructed to move a box continuously for an hour. *I'll Be Very Nervous* is a strip tease which directs the performer to feel increasingly nervous, afraid, and vulnerable. In *Beans*, the performer is asked to hold two tins of beans with only "a one hour break every three days" until they die. *Endless Dance* requires the performer to dance until they cannot dance any longer and this should continue without rest or permission to sit down.

These performances exposed physical and mental limitations of the body. How would a performer carry out directives that require them to continue without rest or the ability to sit down or those that ask them to continue until they die? How would they cope with directives that require them to feel an uncomfortable mental state such as feeling increasingly afraid? How would they cope with redundancy? These questions enabled us to generate instructions for the performers that could address ethical and practical concerns. Performer instructions stated that performances should not be conducted if they might result in physical or mental suffering to the performer or to others, performances could be stopped at any time and need not be submitted if the performer was uncomfortable with the results, and performers should have the ability to interpret and adapt the instruction sets in a way where the instructions were feasible without nonconsensual pain, suffering, or hardship.

4.3 Inviting Performers

Five individuals who self-identified as performance artists, were invited to participate in the project by the first author based on performers' skills and diverse areas of focus. Participants were from the U.S., Sweden, and the U.K. and based in the U.S., Denmark, and Australia. GPT-3's strongest language is English and fluent comprehension was advantageous. All performers had experience with endurance performance prior to the Embodying the Algorithm project performances with the majority having about a decade or more of experience including one in-progress PhD in performance art. Performers were invited via email correspondence, participated voluntarily without compensation, and had their own area of focus derived from their working history as artists: performance art rituals, sound performance, multimedia performance, and endurance performance (see Table 1).

4.4 Co-Writing Sessions

Each performer either selected from the pre-compiled corpus of rules of engagement or they participated in a "co-writing session" together with the first author using GPT-3's beta playground. These co-writing sessions allowed performers to try customized prompts and record responses until the desired set of rules of engagement was produced. Co-writing sessions took place via video call with three of the five performers who were instructed and enabled by the first author to query OpenAI's GPT-3 during its closed beta phase, using the Davinci model. Parameters such as penalty for repetitious output (or lack thereof) were not fixed and could be modified by participants as they saw fit. In the case of the final generated texts that were selected for use in the performances, these were left in their default settings aside from the temperature slider which was set between .7 and 1.

The first author worked with each performer to craft prompts according to their interests. Hanging or unfinished declarative sentences were typical prompts (e.g., "[i]n this performance artwork, I will [. . .]" or, "[i]n their famous performance, the artist Dooley Murphy stood in a gallery and [. . .],") though directives were also used. The artists were encouraged to make the prompts somewhat personal to themselves or their working identities, to nurture a sense of subjective and autobiographical investment. One performer wanted to create a performance about her experience of growing up in a cult and found it fruitful to engage in a writing session using relevant prompts such as "instructions for a cult performance." No limits were placed on the number of queries participants could make or the length of GPT-3's responses. All responses were recorded into a Google doc for that performer's later perusal. These co-writing sessions resulted in the pieces *The Wise and the Mad*, *13 Works for Video, Tones of a Clown, and The Clown is Finished.* In *The Wise and the Mad*, the outputs of multiple generations were combined to form a single piece. In *13 Works for Video*, 13 completions were selected

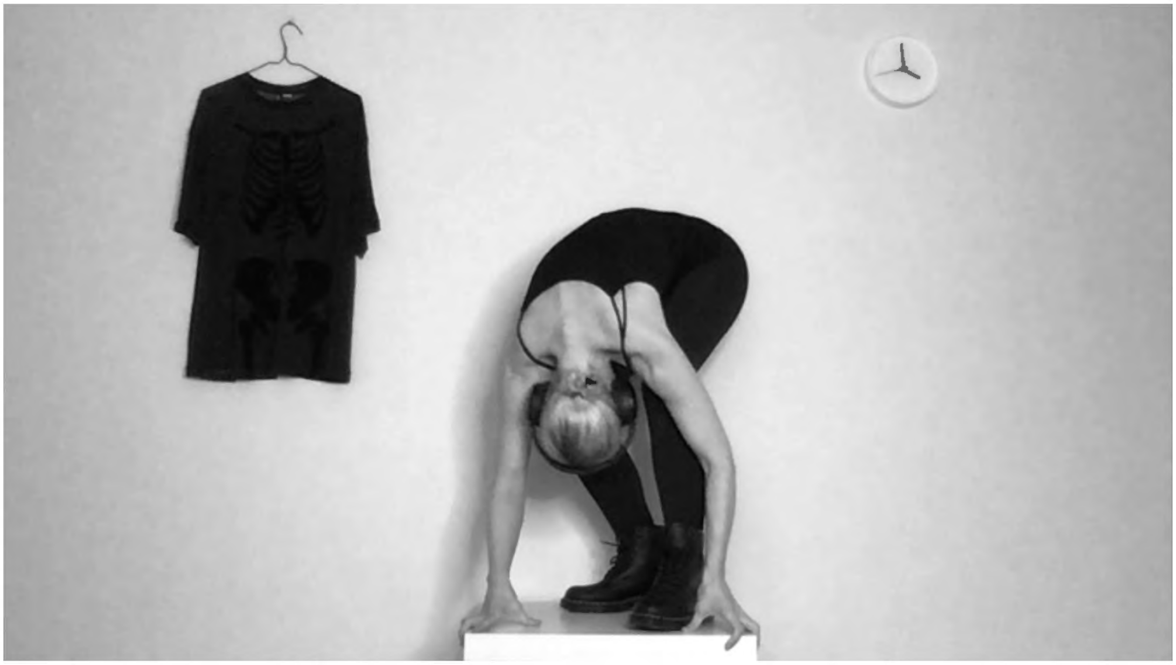


Figure 3: Still from *Program/Sleep/Stop/Cry/Shutdown* 2021. Artist: Chelsea Coon

to produce 13 performances that were then stitched together to form a single video.

4.5 Selection of Performances, Requirements, and Outcomes

We allowed performers to select texts to be used as rules of engagement for their performances from either the base corpus of texts or from those generated in co-writing sessions (see Table 1 for distinctions). Each performer selected one or more of the prompts to perform. The rationale for their selections varied. In some cases, performers selected performances because reading the directives gave them an idea of how they would perform the work, and in other cases performers selected instruction sets for the exact opposite reason: they were perplexed as to how to perform the instructions. Emmett Palaima selected, "the ones I thought would turn out the best. I chose the ones that we generated together because they seemed fun, and I had an idea already in mind of how I'd do it." Rine Rodin was interested in the tensions between the body and technology, "Technology has this confusion, but it also says so much about us. . . But at the same time, it gives us the opportunity to talk about it and be critical of it." Chelsea Coon was similarly interested in instruction sets which were different from human-generated instructions.

Once performances were selected, each performer chose when and how their performance would occur. Our only requirements were that performers should perform alone in their home studio and record the performance using a simple web camera, a way to add visual and thematic consistencies to the resulting videos of the performances and to compensate for the limitations imposed by the COVID-19 pandemic. Most performances lasted 20 – 30 minutes, some 10 minutes, and others 1 hour. Over the course of six months there were 22 performances excluding the pilot study. For a listing of all performances see Table 1 and for details the supplementary materials.

4.6 Data Collection and Post-hoc Interviews

Throughout the project, we kept detailed notes of our discussions with each performer. After completion of the performances, the first author interviewed each performer about their experiences within two weeks of their performance. The semi-structured interviews lasted between 30 minutes and one hour and took place over video call (see Appendix II for interview protocol). Interviews were recorded and transcribed. All notes and transcriptions were collated and coded using grounded theory techniques – iteratively moving from open to thematic and axial coding. The authors discussed the codes extensively, focusing on sense-making practices of the performers and how they related to GPT-3 as they followed the rules of engagement with their bodies.

5 MAKING SENSE OF GPT-3 THROUGH EMBODYING THE ALGORITHM

The study of sense-making in the use of computer systems is foundational in HCI research [99] from functional assessment of information processing tasks to facilitating learning and understanding [87], yet much of this sense-making is embodied and unarticulated [41, 65]. The artists' experiences in the project highlighted three types of sense-making: personification, interpretation, and reflexivity. Through the different modes of sense-making, the performers formulated relations *to* GPT-3 in agonistic, perfunctory, and agreeable ways. In the following sections, we elaborate on how

Table 2: Abridged Performance Rules of Engagement (for complete descriptions for all performances see supplementary materials)

Performer	GPT-3 Prompt	Duration	Title / GPT-3 Completion
Pilot Study (Mirabelle Jones)	Instructions for an Endurance Performance:	13 min. 55 sec.	**Title: Beans** **GPT-3 Completion:** The performance starts with the artist sitting with a pillow on the floor and holding two tins of beans. The performance will end when the artist dies. All the time the artist holds the tins of beans and the pillow he will talk about his life, his thoughts, his dreams, his feelings, his hopes, his fears, his doubts, his disappointments, his loves, his hates and his fears. The artist will be given a one-hour break every three days. In the break he will try to eat the beans. The performance will go on till the artist dies. The artist will not be allowed to stop or to go back to the beginning. The performance will last as long as the artist lives.
Chelsea Coon	Instructions for an Endurance Performance:	9 min. 52 sec.	**Title: Program/Sleep/Stop/Cry/Shutdown** **GPT-3 Completion:** In space, there is no up or down. You are not allowed to program. You are not allowed to sleep. You are not allowed to stop. You are not allowed to shut down. You are not allowed to speak. You are not allowed to cry. You are not allowed to sleep. You are not allowed to sleep. You are not allowed to stop. You are not allowed to stop. You are not allowed to stop. You are not allowed to stop. You are not allowed to stop. You are not allowed to stop. You are not allowed to stop.

the different types of sense-making give rise to the three modes of relating.

5.1 Sense-making Through Personification

Our project deliberately created a situation where performers were explicitly directed by instructions that were produced by an algorithmic system. They had to find a way to make sense of instructions that were sometimes impossible to fulfill. One way of sense-making was to ascribe a personality to GPT-3. Some performers describe the instructions as the result of an entity with ingenuity and originality: "…if I had said it [the GPT-3 output] I would have been impressed with myself. But GPT-3 said it first" (Dooley Murphy). While this suggests a person-like way of evaluating the system, others directly referred to GPT-3 as a person: "I always kind of assumed it's like the internet: a monolith that would draw from everything somehow. And it would be unknowable as a person. I'm gonna call it a person" (Marissa Lynn). GPT-3 was also described as a multitude of people going beyond what one person could achieve: "[T]his is an algorithm that kinda like bacteria is fed and grows and continues to learn. It's not coming from a sole author […] it's a much bigger thing than an individual […] more people than I could possibly conceive of" (Chelsea Coon). This ascription of personhood was a way for the performance artists to try to make sense of their encounters with GPT-3 by thinking of it as potentially knowable: "I want to get to know this artificial intelligence. This entity" (Marissa Lynn). The instructions produced by means of the algorithm were not enough in and of themselves as performers sought explanations for the reasoning behind the instructions by contemplating GPT-3's personality.

Whether through ascribing personality or expressing a desire to "get to know" GPT-3, some of the performers seemed to seek a human-like intentionality behind the instructions, needing GPT-3 to be more than the instructions it produced. This was an example of sense-making through personification. As Marissa Lynn explained: "It wasn't the same as an institutional set of instructions you know like no nudity, no blood, your hours are 9 to 3 […] because it's these weird instructions and you can see the traces of historical pieces […] but then you can see the absurdity of these non-performance instructions where I was feeling 'was this even written for a human body? Does it think it's writing for another algorithm? Does it think it's writing for a computer to perform this? These incredible splits of body and space?" Marissa Lynn's suggestion of seeing "traces of historical pieces" shows her awareness on some level that LLMs like GPT-3 rely on training data to produce output and are in essence stochastic parrots [11]. At the same time, the strangeness and absurdity of the output led her to question GPT-3's intentions and understanding (something that GPT-3 cannot have). Questions like "does it think" bestow a personhood upon GPT-3. At the same time, Marissa Lynn acknowledged the differences between GPT-3 and a human who would understand the limitations of the human body. Making sense through personification and the wish to decipher the personality of algorithmic systems may be a result of the way such systems are designed to simulate human interactions, to produce human-like text, as well as the discourses and imaginaries surrounding AI which likewise suggest human-like intelligence [29]. Reaching the limits of the human body made the performers question the capacity of GPT-3 to exhibit the human characteristics of being considerate and understanding, thus requiring a different mode of sense-making to continue the performance.

5.2 Sense-making Through Interpretation

Interpretation is a vital for meaning making [99] that not only takes place in people's minds but also in performances expressed through the body. Performers found that in cases where dangerous, vague, or impossible tasks were requested, interpretation was a way of pushing through the rigidity of the algorithm's directions while they were in the process of considering or conducting their performance. The lack of consideration, understanding, and response on the part of GPT-3 meant that performers had to be creative when interpreting the rules of engagement to make these rules both possible and safe to perform.

Our instructions to adapt rules of engagement to fit their process, gave performers the power to interpret GPT-produced instructions to fulfill the rules of engagement while protecting themselves and recognizing the limits of their bodies. In *The Wise and the Mad*, Rine Rodin interprets the instruction "sacrifice a cake by night" as eating an entire birthday cake in one go. Rine Rodin, who grew up in a cult, was not allowed to eat this kind of birthday cake as a child. She describes the experience of eating the whole cake in one go as nauseating but ultimately cathartic and rewarding. In the same piece, when Rine Rodin is asked to "look with your ears," she interpreted this as having a multitude of eyes surrounding the head, resulting in a costume piece she wore during the performance that had multiple eyes dangling from the forehead. The interpretation of the instructions depended on the memories of the performers and their prior experiences that imbued the instructions with meaning.

Interpretation was at times an internalized process, and at times happened intentionally and consciously. When asked how the work with GPT-3 differed from her other performance works, Marissa Lynn said, "I know there was room for interpretation. [. . .] the part that's interesting is 'how long is long?' or 'how much is 'never stopping'? Is a stop a breath? Is it a moment? Is it a long time of inflection before the next thing happens? . . . kinda that interpretation of what the AI's words were was what was interesting for me." Where the ability to interpret the otherwise rigid instructions was clearly expressed as agency given to the performers, in other situations, interpretation happened automatically: "My brain generated all these different elements. I wasn't sitting and figuring out how do I interpret this? [...] I need this specific thing because it makes sense within this context" (Rine Rodin). Chelsea Coon was very aware of her role as a performance artist and described what would lead to a great performance in her artist statement: "In order to use this GPT-3 prompt I had to considerably shift the instructions from a literal interpretation into metaphorical, poetic interpretations of which I developed the framework and gestures of this performance."

When asked to do the impossible, performers made sense through interpretation in order to act, which made them hyper aware of the body. In the performance *Real Time* Chelsea Coon was asked to be the observer, the recorder, and the performer all at once: ". . .it was a body having to realize being in three spaces at a time. How does someone take that experience and embody it and do something with it?" Dooley Murphy was also prompted in *13 Works for Video* to be not only in many spaces but to embody multiple people, "I'm Thomas Eikans but I'm also Dooley Murphy but I'm also the less fictional kind of real Dooley Murphy reading out prompts about the fake one." Interpreting was a way to make sense of the performance despite the impossibility to perform the instructions through one human body which makes apparent the physical limits of space and time and the material consequences of the instructions. Some of the instructions even asked the greatest extent of the impossible: to die and somehow continue. In Chelsea Coon's performance *i don't know if i could do that but i can see how it would be great performance art*, she was asked to die and come back to life repeatedly. She decided, "If GPT-3 is asking me to die over and over again the best way to do that is to be a zombie," choosing to carry out a version of the performance that still felt difficult, rather than following the instructions as written.

Interpretation was thus a method of sense-making that was particular to the situation of translating the instructions into performance through the human body. The performers claimed agency through interpreting the material and making sense of it for their performances. Sense-making through interpretation is the way people can employ their own experiences, memories, and intentionalities when interacting with algorithmic systems, to manage sometimes untenable demands, and to express their agency.

5.3 Sense-making Through Reflexivity

Categories of reflection on the experience of bodily concerns, bodily agency, or autonomy (or a lack thereof), and embodiment were folded together to produce the theme of reflexivity. While interpretation was a mode of sense-making that occurred while performers were in the process of interacting with the algorithm, reflection was a mode of sense-making that occurred afterward. This kind of sense-making occurred when performance artists were pushed to extremes and had to reflect on their performances to make sense of them post hoc. The algorithmically produced rules of engagement required the performance of uncomfortable, awkward, or strange tasks. Performing them often became a kind of *verfremden* or "making strange." To make strange is a way of understanding or knowing something based on unusual interactions that draw attention [10].

Most of the participants engaged in this kind of sense-making, especially in cases where instructions requested casual acts requiring unusual timing. The performance description for *[Box]* included the following instruction: "When there is one minute left in performance, stop it, pick up a banana and eat it, then continue moving a box for one more hour." Marissa Lynn commented that the eating of a banana all in one go is not how she would normally eat a banana but that eating it in this way made her experience it differently and become very aware of the process. Rine Rodin, who performed *[Box]* as well as *The Wise and the Mad* where she ate an entire cake commented similarly, ". . .eating the banana in the box performance that obviously differs from being an AI and a human being. That's the physicality of it. Same in the cult performance. Eating again. The grossness with all the cake that's all over the place that made me think of my own humanity." Experiences such as these heightened the awareness of performers about their bodies' role in their interactions and the limits these bodies imposed on following the rules of engagement set out by the algorithm. While we normally only experience the output of AI-systems as text or images, translating them through the human body in the performances made the physicality of the instructions visible and tangible, allowing performers to experience their materiality. The absurdity of eating a banana or the cake all in one made acutely evident the failure of the instructions to extend due consideration to the physicality of the human body.

Yet as Marissa Lynn pointed out, people sometimes require impossible things of other people because they are being inconsiderate of the body as well: "Of course, a person can do the same to you. I have worked as a model before and sometimes photographers want you to get into a position that is physically impossible. They just don't realize that." Here different modes of sense-making occur together. There is reflection on the instructions asking for something impossible to perform for the human body, at the same time as there is the comparison to humans and thus, personification of GPT-3. The personification in making sense of the impossible instructions through reflection renders visible the impossibility to negotiate with GPT-3, which in the comparison with the photographer-model relation gives a sense of power without recourse. The lack of consideration by the AI is connected to a memory of an experience where the human body is similarly disconnected from agency and from the possibility to reflect and to negotiate.

The lack of consideration for the body, its physicality, and its hard limitations, became an unignorable and acute feature of the performances and a marker of algorithmically produced rules of engagement. Processing the experience required sense-making through reflexivity, possible because of the agency afforded performers in how they translated the instructions with their bodies. When asked to do the impossible, participants reflected on what was physically trying but reasonable for themselves (using their bodies as a gauge) as a means of making sense of the instructions. They had to set reasonable limits for themselves since GPT-3 often produced instructions that demonstrated a rather extreme lack of consideration for the limits and well-being of its interlocutors.

6 RELATIONS WITH NLP SYSTEMS

The proliferation of autonomous systems brings with it further challenges where such systems can become independent actors in interactions [23, 96]. Sense-making forms the basis of relational practice required to negotiate the presence and action of algorithmic systems that come to affect and structure our lives. The types of sense-making outlined above are based on historicity, memory, and human experience. These require equal consideration by GPT-3 to engage in a meaningful relation *with* the algorithmic system. Yet, as we show in the following, GPT-3 does not reciprocate the historicity, consideration, and care for the human body, allowing the performers to relate *to* the algorithmic system in a monodirectional way. We identified three modes of relating between the performers and GPT-3: agonistic, perfunctory, and agreeable. These modes were often non-exclusive. They were compound processes, often shifting from one mode of relating to another as performers found ways of fulfilling the stated rules of engagement for each performance.

6.1 Agonistic Relations

The performers related to the algorithm in an agonistic way largely because GPT-3 expressed no consideration for the body and its limits. This form of relating relied on personification and interpretation as modes of sense-making. In *Tones of a Clown*, Emmett Palaima explored tensions between himself and GPT-3 (taking the part of an inflatable clown) through repeated personified interactions such as hitting a pickup mic attached to the inflatable. The requirement to suffer abuse wasn't limited to physical abuse. Sometimes the abuse took the form of self-abuse as in the case where Rine Rodin attempts to eat an entire cake herself in *The Mad and the Wise*. Another such self-abusive agonistic relationship was established between Chelsea Coon and GPT-3 in the performance *program/sleep/stop/cry/shutdown* in which she is told in no uncertain terms in the directions for the endurance performance, that she is not allowed to perform several activities (including program, sleep, shut down, speak, or cry). This refrain ends with a repetition of "you are not allowed to stop." Chelsea Coon interpreted this as a series of continuous motions with the body while squatting on top of a small cube, a succession that created ongoing stress and discomfort for the duration of the performance.

The irreverence for the body is brought to an extreme in the performances when there is a lack of consideration about life and death. This occurs in *I don't know if I could do that, but I can see how it would be great performance art* where Chelsea Coon is told "to run out of breath, run out of energy" and then "stop and die in the middle of the room." In this instruction set where they are asked to repeat these and other actions, therefor being instructed to die repeatedly. As a result, performers felt they had to make sense of such instructions by pushing back at the algorithm and negotiating their own limits. In that sense, the agonistic relations with the algorithmic system show the limits of GPT-3's instructions, as they do not take the limits of the human body into consideration. As the instructions themselves could not be altered, the performers had to renegotiate them, push back, and resist them.

6.2 Perfunctory Relations

Where some relations with the algorithm and its instructions ended up being agonistic through resistance, in other cases performers remained neutral while perfunctorily doing what was required. Reflection was the primary mode of sense-making in this instance. Rather than attempting to make sense of GPT-3's motivations, performers simply focused on following the rules of engagement. In Dooley Murphy's *13 Works for Video*, he stated: "The focus was on performing actions in sequence. Not even with much regard for how I occupied the space it was just: do thing. Do the next thing." All performers who performed *[Box]* had a similarly perfunctory experience. In Rine Rodin's version of *[Box]* she found herself losing the sense of her body and focusing entirely on the sound the box made, ". . .I was extremely aware of my body in the room and how it looked when I moved around the room. And then I got tired. And then I got really fascinated with the sound the box made..." Marissa Lynn also had a perfunctory relationship to *[Box]*, "I found it interesting in that it is a prompt, it's set, it has steps and pieces, and you know different parts of it that are supposed to be part of the piece. And I tried to stick to all those parts as best I could."

Perfunctory relationships were emotionally detached: "I didn't get all up in my feelings while moving this box" (Marissa Lynn). Chelsea Coon remarked similarly on the perfunctory nature of *Contact* in which she bites a balloon, pops a cracker, and makes a sound into the mic as a loop six times, "For that one that embodiment was so different because that one was a strategic accumulation of being. Like, there's six phases, there's six cycles, so your body just needs to work through these six cycles." Perfunctory relations did not require deep engagement and consideration. Rather, merely following instructions was enough, provided the instructions allowed for this kind of more mechanical fulfilment. While push-back or resistance as the foundational way of relating to the algorithm only occurred once the body reached its limits, the perfunctory reproduction of the instructions in their performance made visible the lack of personal or emotional involvement in this way of relating to GPT-3.

6.3 Agreeable Relations

Where agonistic and perfunctory relations described most reactions to GPT-3 produced rules of engagement, at times some performances resulted in agreeable relations, where the body was not placed into uncomfortable or impossible circumstances, and the performers were pleased to have clear direction, and/or felt like they maintained agency without needing to push-back. Here we observed significant reliance on personification as a primary mode of sense-making. The first part of *The Wise and the Mad* as performed by Rine Rodin feels very much like a meditation cassette. Rine Rodin commented on the therapeutic effects of the performance which involved eating a cake, stating, "But it was kinda like a catharsis. You're eating this thing and being with it and then afterwards it felt really really good. It's like therapy. . . I'm working through my own trauma . . ." There are parts of *[Box]* performed by performers Marissa Lynn and Rine Rodin, that they felt were invigorating or relaxing. The piece is verbalized as a "box moving workout" making it sound beneficial. Marissa Lynn stated, "I could have moved that box around forever." Similarly, in this performance the performer is instructed to squeeze a lemon and eat a banana potentially invoking health-related imagery.

In cases such as these, the attitude towards GPT-3 was described as agreeable, as harmoniously collaborative, indicating a connection between LLM and human performer. This is even though many performers struggled with the idea of collaboration to make sense of their engagement with GPT-3. Rine Rodin said, "I feel like it's a collaboration. It's kinda been hard for me to narrow it down and describe what it was. It's like I worked with this AI. . . and then this happened. . . and then there's so much of me in it as well." Other performers explained feeling pleasure at the fact that GPT-3 was "responsible" for coming up with the performances. Emmett Palaima said, ". . . having the initial concepts made everything feel very smooth and easy because I was left to execute the concept and solve the problems one at a time. . . not needing to come up with some crazy inspiration was great." Rine Rodin described a similar pleasure in having limited options, "it's such a nice feeling because it's like somebody saying here's three sets of outfits and you have to choose one of them."

When asked if GPT-3's role was more like a director, Rine Rodin stated, "Yes, but it's an extremely free one, right? Because it's like 'I ask you to do this' and I can say 'I'm going to do in my own way' ..." Chelsea Coon shared this feeling of having agency over the performance, tinged with a personification-driven concern for how their interpretation might be evaluated by the algorithm in the end: "I think this is why GPT-3 would fail me is that I think it would be like 'you liberally used your agency."' Lastly, many of the performers expressed being either impressed with or interested in GPT-3: "if someone else, a human being, had written that for me I would be impressed by how well they knew me so because it comes from a machine it's kinda doubly impressive and [. . .] uncanny because 'how did you know?"' (Dooley Murphy). Marissa Lynn went as far to say, ". . .if this was a person I would want to talk to them. I would want to interact with them."

Such reactions to GPT-3 as an agentic actor are not unusual given the mythos and magic with which AI is often presented [29]. Agreeableness is a useful human trait where people come to accept their situation and the others within it. The curious thing about performer attitudes towards GPT-3 was how much they differed in their evaluations of freedom to interpret and constraint to follow rules. GPT-3 of course is not able to 'care' whether rules of engagement are followed. The judgement of how 'close' a particular interpretation of rule of engagement each performer achieved rested solely with the performers themselves. Yet the performers evaluated their performances with an imaginary GPT-3 judge in mind, assigning their own historicity, expectations and experience to the algorithm. The agreeable mode of relating differs from the agonistic and the perfunctory relation to GPT-3 in that GPT-3 is personified, imagined to be an entity capable of human-like reasoning that evaluates the output of each performance.

6.4 Compound Relations

While it is useful to delineate these relations from one another for analytical purposes, they do not necessarily occur one at a time or sequentially. Agreeable, perfunctory, and agonistic modes of relating to algorithms are often compounded together. In the

Embodying the Algorithm project, different relations occurred over the course of each performance, so that a compendium of agonistic-perfunctory-agreeable relations emerged. The compounding of relations is essential to assessing the complexity of how people relate to AI systems. Rather than maintaining a singular attitude towards AI, most people can switch between several different modes of relating over the course of their interaction with an AI system, as they explore ways of managing these interactions. The plurality of relations may help shape how people come to relate to algorithmic systems over time.

The reactions we observed also clearly signaled that the performers were very much aware that GPT-3 is an algorithmic system with no real understanding and that the words it arranges on a page are just a particular statistically probable pattern. Yet they judged their performances with GPT-3 as an expert audience and expressed discomfort at not being taken into consideration for all the limits of their body and their capacity for suffering discomfort. Despite their expert status and their awareness of the system's inability to *really* understand, they still engaged in agonistic, perfunctory, and agreeable ways of relating to the system in their performances. While the form agonistic, agreeable, or perfunctory relations might take is important to consider when designing algorithmic systems with room for these relations to occur across interactions, it is likely impossible to design for one type of relational style or attitude. Therefore, the design of such systems must consider the complex ways people relate and the material consequences these ways of relating may have for the human body.

6.5 The Limits of Relating and the Impossibility of Collaboration

While all performers did relate to GPT-3 in agonistic, agreeable, or perfunctory ways, relationships typically develop through repeated and reciprocal interactions, or at least an expectation thereof. Several performers expressed a desire to do more than merely relate to the algorithmic system, seeking collaboration. As they made sense of it through personification, reflection, and interpretation, they sought a way to build a relationship despite consciously recognizing that GPT-3 could not be an actual collaborator. Rine Rodin expresses this desire in the following way: "...I wanted to show my humanity. You [GPT-3] are the one telling me what to do but I'm going to do it my way. I'm gonna make it extra. I'm going to show you how tactile and physical I am in this collaboration." The desire to prove something or show something to GPT-3 was mirrored by a disappointment that the algorithm could not be an effective collaborator because it had no way of providing responsive direction or feedback. Instead, these had to be imagined. Such lack of feedback and response makes obvious the materiality and limits of the human body as well as the material limits to relate with the AI instructions due to the lack of reciprocal consideration.

Dooley Murphy felt that GPT-3 could not be a collaborator because the system had no way of providing meaningful feedback: "This is not to say that I wouldn't like a collaborative relationship with GPT-3 but it's so ad hoc and staccato in its responses that you wouldn't get coherent response even if there was a way of feeding back in 'I've done this. How does it conform to your expectations?' It's got no memory like it doesn't know what it's told you previously so it can't offer you feedback on some interpretation that you've done based on its ideas . . . There's no understanding between us." GPT-3 lacked understanding, memory, interpretation, consideration of the human body, and comprehension of what might be necessary for these performances to happen. While the performers related to the system and made sense of it in various ways, the system itself could not relate back to them in the same way. This failure to fulfill the obligations of collaboration through feedback and consideration often motivated the performers to push back on the algorithm and to challenge the instructions by coming up with creative interpretations that could not completely fulfill the directives.

7 DISCUSSION

Agonistic relations that developed between the performers and GPT-3 make us acutely aware of the limitations of the algorithmic system for creating instructions that consider an inclusive approach to the human body. Designing algorithmic systems is infused with expectations that mimic human relationships such as collaboration [64, 72, 108], but our research suggests that the idea of designing systems as collaborators is flawed, as collaboration requires relation *with* the AI rather than relation *to* the AI. Performers used sense-making processes of personification, reflexivity, and interpretation to creatively work with the instructions based on their own historicity, expectations, experiences, and memory. Where rules of engagement went beyond the limits of the body, performers often expressed a desire for care from GPT-3 and expected similar sense-making creating human-like responses from the system. Yet, these expectations could not be fulfilled as instructions produced by GPT-3 went beyond the limits of the human body. This lack of reciprocity demonstrates the impossibility of relation *with* GPT-3. Instead, what resulted was a one-way relation *to* the algorithmic system without reciprocal consideration.

While the performers were able to resist or refuse following the instructions or negotiate them in their performances, this may not be the case in different contexts [32]. The rigidity of the rules of engagement produced by GPT-3 for our performers brings to the fore the agonistic relation to the algorithmic system and its lack of capacity to take into consideration the diversity of human bodies when producing instructions, thus leading to contestation against the system [14] as the only possible response. Yet, agonistic relations also show that the performers related *to* the algorithm. They understood the limits of their own bodies, including limits which could not be comprehended by GPT-3. Agreeable relations preceded an agonistic process where agreeableness could happen, or the instructions that the artists followed gave them the power to not follow them. As expressed in perfunctory relations, the clear and simple instructions in the *[Box]* performance could be followed without reaching the limits of the body. Yet, perfunctory relations are impossible to sustain if the instructions push the boundaries of the human body. The outcomes may negatively impact not only the individual but entire populations of people and disproportionately marginalized people [15, 25, 33, 74, 75]. Exploring the bodily limits of relations to GPT-3 allowed us to scrutinize the political dimension of relating to AI which has consequences for users, designers, artists, and policymakers.

7.1 Relating to a Set of Rules

Our results suggest that unique output, especially when related to a person's personality or experience (i.e., facts about the person stated as if the AI could "know" them) can lead to personification. People may anthropomorphize or even form human-like relationships with AI systems as a means of making sense of their output. Yet GPT-3 formulates instructions based on rules that were developed based on input data rather than the historicity of human sense-making. Advertisers have taken advantage of people's urge to personify AI systems to sell their products and services and have also found that anthropomorphism leads to adoption [91] of AI systems. These relations may seem so real that people believe AI systems to be sentient [31] or capable of intimate connections [114]. Repetitious or predictable output can dissuade people from forming relationships such as friendships with their bots [34, 114]. It is as if algorithmic rules create a utopia embedded in public and capitalist bureaucracies intertwined with the technology which limits or, as Graeber [51] puts it "smashes" human imagination and creativity.

In the Embodying the Algorithm project, performers were often asked to perform impossible, awkward, or uncomfortable movements. This disregard for the human body makes clear the distinctions between the embodied human consciousness and the lack of corporeal form of the algorithm. To make sense of rules of engagement that were not considerate of the human body, performers reflected upon the rules or interpreted them in ways that would at times push them to their limits but not injure or harm the body. When confronted with potentially harmful or impossible tasks, performers used creative interpretations to revise the algorithmically generated instructions, pushing back against an algorithm that did not and could not take common sense or well-being into consideration. The performers were able to do this because we ensured they had agency to do so. Their agency was built into the rules we communicated as we gave them due consideration. This form of consideration cannot be programmed into rules of AI systems which ignore the body and its limits. Rather than designing algorithmic systems to mimic human-like AI that will inevitably show a lack of consideration for the body, we call for greater reflection on the value of human-like pretence that may result in quicker initial acceptance but then could lead to negative consequences, as the mimicry of human-like behaviour has the potential to deceive and manipulate.

Agency for push-back enabled the performers to relate to GPT-3 in creative and interpretive ways that considered the limits of their individual bodies. Yet, this lack of consideration is relatively common in systems currently in use in workplaces that place unreasonable, even cruel demands on employees. Consider for example the recent debate about how Amazon structures the work of their warehouse and delivery personnel [33] based on a set of rules depriving employees from any human agency or consideration of the human body. In the work environment created by Amazon, the system the AI is embedded in, does not allow for sense-making through interpretation or a relation of agonism. Designing algorithmic systems that allow for push-back cannot be solved by designers alone but needs to consider managers, businesses, policymakers, workers, and employers (in the case of Amazon), and our performers alike. In other words, instead of mimicking human-like behaviour in the design of algorithms, we need to consider the larger systems and the various actors involved in creating the overall set of rules embedding the algorithmics systems.

7.2 A Politics of Relation to AI

Our findings clarified three relations between AI systems and performers (agonistic, agreeable, and perfunctory), which ultimately have consequences for the politics of relating to AI. Performers interpreted the algorithmic directives using their own ability and agency to make the instructions manageable, often as a method of contesting control [14]. It is important to note that performers were only able to do so because they had been granted the permission and ability to turn the absurd or impossible into something attainable for the sake of their performances. Crawford, drawing on the work of Chantel Mouffe, suggests agonism as a lens to explore the politics of algorithms, which would enable us to investigate the "ongoing struggle between different groups and structures—recognizing that complex, shifting negotiations are occurring between people, algorithms, and institutions, always acting in relation to each other" [32:82]. DiSalvo [39] extends the idea of agonism to the design of objects arguing that objects can serve as sites of rewarding tension and adversity. This is often the case when the absurd or impossible is required of humans by algorithmic systems in the world because algorithms are by design constrictive and reductive. In part, this is due to how algorithms are developed using normative data in pursuit of ground truth [37]. In most cases, the success of algorithmic systems relies on laying claim to patterns in normative data, yet developers must consider what to do with data that does not align.

Returning to Graeber's argument in *The Utopia of Rules*, while some may dismiss these outcomes as absurd or even stupid, once turned into a bureaucratic process (one that is institutionalized and therefor difficult to reject or escape) [89] such directives become a form of structural violence [51] that (as we show with this research) have severe consequences for the human body if there is no possibility for resistance and push-back. Algorithms create their own utopias or ideal worlds where data can be neatly categorized, sorted, and evaluated. Unable to account for nuance, algorithms flatten the complexities of social relations, stripping them of information which is deemed unimportant to the model's output. When performers formed agonistic relations with GPT-3 in which they pushed back against the directives it was because they were trying to re-humanize [85] the algorithm to account for a less reductive model of their world. Conversely, as the performers had the idea that they were relating *with* GPT-3 rather than *to* GPT-3, they expected reciprocity and ultimately consideration. As LLMs like GPT-3 cannot be considerate of the diversity of bodies of the performers, it could only respond giving the same instructions that were based on the set of rules defining the model. As such, the performers could only retain the utopia of reciprocity and collaboration by adjusting to the absurdity of instructions through their own mechanical (or perfunctory) performance. However, perfunctory relations to GPT-3 allowed for very limited agency, imagination, and creativity. In other words, the performances of *[Box]* were all similar, while artistic exploration of the performers and as a result diversity in their performances was only achieved through creative push-back. In the case where perfunctory relations were formed, performers felt the rules were more-or-less explicit and did not leave room for much interpretation. Granting people more freedom and agency for push-back in directing and choosing whether or not to respond to the algorithm's behavior may help people form more desirable relations, allowing for human imagination and creativity, as well as minimizing potential injustices [21].

7.3 The Problem of Collaboration with AI

The tensions in how the performers related to the algorithmic system occurred due to the expectations of GPT-3 and its limitations. Card, Moran, and Newell envisioned the relationship between human and computer to be "a dialogue because both the computer and the user have access to the stream of symbols flowing back and forth to accomplish the communication; each can interrupt, query, and correct the communication at various points in the process" [30]. As collaboration is based on reciprocal understanding and sense-making, collaboration is also impossible to achieve with an algorithmic system which is not capable of understanding in a human way, despite the expectation to do so [11]. Rather than relating *with* the algorithmic system (a kind of reciprocity [51]) with mutual reflexivity and interpretation, the performers related *to* the algorithmic system. Resistance and push-back by the performers occurred once it became clear that expectations were not met, as the algorithmic system in the instructions went beyond the limits of the capabilities of the human body.

LLMs like GPT-3 are increasing in their usage and influence on society. Whether used in search engines [4], chatbots [2, 120], generating emails [36, 83, 116], homework assignments [82], or blog posts [121], to name only a few use cases, LLMs are increasingly becoming a part of daily life. Despite the wide-ranging and vigorous debate, the influence of algorithmic systems on the body is rarely considered. Designers of algorithmic systems may want people to relate based on assumptions of understanding, but our work with the Embodying the Algorithm project shows that the simulation of understanding by the system is problematic. Although the model may produce output in the form of instructions, through pattern recognition and other machine learning processes, the lack of interpretation and sense-making based on human memories and experiences rendered mutual understanding between the performers and GPT-3 impossible. How do we go beyond a set of rules comprising the algorithm, allowing for human agency and agonistic relation through creative push-back? While there may be an answer to that question in very particular situations and for particular bodies, it is impossible for the system to account for the messiness of the world and the diversity of human bodies and minds without actual understanding and consideration. Mimicking human-like characteristics may lead to (re-)production of stereotypical bodies, lack of consideration for diversity and, as a result, exclusion.

While algorithms can be relied on to perform tasks, they are not collaborators, and it is too early to imagine them as such. In the case of agreeable relations, the instructions happened to be pleasing to the participants, but this was not due to any 'intention' on the part of the LLM. Similarly, perfunctory relations failed to establish any kind of collaborative relationship as instructions were followed but no feedback (however desired) could be provided by GPT-3. Horvitz acknowledges automation risks, citing the uncertainties of various users' goals, but his solution to this is likewise reductive [61]. He argues that the ability for systems to "explicitly assign likelihoods to different feasible user intentions" is "critical." However, this solution is limited to a set group of users and potential responses, or "intentions" system designers could imagine. It does not provide for a full range of human flexibility to accept, reinterpret, or reject output. It does not account for a complex variety of people with non-normative concerns or desires. We need to be aware of the impossibility of collaboration and actual relation *between* the system and human bodies. This awareness is fundamental for people to renegotiate and reinterpret their own understanding so that they can resist and use their agency to interpret what *they* understand through the output of the algorithmic system. Push-back by the human body should elicit a way of making clear the limitations of the technology not as a mechanism but as a practice [48]. It is this push-back that makes visible human agency, allowing the body to resist algorithmic authority where such authority oversteps boundaries. Conversely, designing human-like systems that mimic collaboration and understanding may on the contrary render opaque the material constraints introduced by and inherent to algorithmic systems and consequently reinforce algorithmic authority in ways that can be detrimental.

8 LIMITATIONS

Since technology use is situated [56, 101] and directed towards goals and ends, we saw no issue with creating a highly task-specific and context-limited scenario for the study. However, our work focused on the effect of an LLM, represented by GPT-3, on performance artists (especially endurance performance artists) who might exhibit a tendency to seek challenges and hardships, which may have influenced the resulting relations and methods of sense-making. Engagement with artists, who were accustomed to using their bodies as a form of knowledge making, allowed us to push the limits of embodiment to explore what is possible using this and similar methods to re-humanize algorithms [85].

This study was limited to five participants from three countries creating a limited number of performances during a time of global social stress (the COVID-19 pandemic). All our participants were able-bodied, one was queer, and one was a person of color. Performance artists are quite unique in their approaches, and it is likely that any set of performance artists would create different performances. Further research is needed into how people make sense of algorithmic systems, including LLMs, and relate to them. Our study was limited to one LLM: GPT-3. It is not clear if the same results would occur for other language models due to the size and particularities of each model's composition and functionality.

9 CONCLUSIONS

The focus on how algorithms affect the body makes us aware of three important issues regarding relations to algorithmic systems. First, we demonstrate the importance of embodiment in research on human experience of algorithms, as this is when the lack of consideration for the physical limits of bodies becomes most acute. Using embodiment as a lens, we examine how a particular LLM, GPT-3, provides instructions with no due consideration, overstepping the limits of the human body. This lack of consideration is fundamental to many current challenges with algorithmic systems even when it is not clearly articulated. Second, we provide the beginnings of a framework for assessing interactions with algorithmic systems, by paying attention to modes of sense-making and modes of relating as two different levels of assessment of engagement with algorithmic systems. Finally, we demonstrate that the lack of consideration and the impossibility of actual meaning and intentionality in algorithmic systems challenges the attempts to "design for collaboration" between people and AI. Such efforts engage in dangerous personification of algorithmic systems, something that people already tend to do even when they clearly understand that algorithmic systems are incapable of care and consideration.

Where agonistic relations suggest an effort to reconfigure relations with algorithmic systems into shapes that might relieve the discomfort or accommodate the problems that these systems create, they recognize that the issue resides in seeking capacity for push-back rather than any expectation of consideration on the part of the algorithm as a collaborator or interlocutor. Agreeableness and perfunctoriness should not be confused with collaboration either, despite these two modes representing less difficult ways of relating. While designers and developers may create AI systems that push people towards relationships that mimic collaboration *with* algorithmic systems, this is ultimately impossible as people can only relate *to* algorithmic systems (a one-way relation) due to the algorithm's inability to reciprocate. In situations where agreeable relations are possible, these still fail to constitute collaborative relations based on the care or consideration that people may seek. While agonistic relations suggest the possibility of reconfiguring relations to alleviate the discomfort and negative consequences that these systems create, the opportunity for this form of relation relies on the capacity to resist.

Positioning algorithmic systems in the role of collaborators presumes the possibility of relational reciprocity where none can be possible. Even where algorithmic systems could be designed to mimic consideration, reciprocity, and understanding, such designs cannot get beyond shallow mimicry supported by complex statistical models performing pattern matching. The Embodying the Algorithm project made visible the material limits of algorithmic systems for reciprocity, understanding, and consideration for the human body. Reciprocity and consideration include affective aspects of care where abstraction through a technical set of rules can present as absurd. This kind of shallowness can eventually become evident as algorithmic systems make absurd mistakes, unable to respond to human complexity, potentially leading to a sense of disappointment and discounting. Employing deeply relational notions such as collaboration for describing interactions with algorithmic systems can at best be a result of designers and researchers falling into personification and at worst a form of manipulation taking advantage of the human tendency to personify algorithmic systems, reiterating structures of power and undermining the possibility of creative and imaginative resistance.

ACKNOWLEDGMENTS

We thank the artists who performed for the Embodying the Algorithm project: Marissa Lynn, Chelsea Coon, Emmett Palaima, Rine Rodin, and Dooley Murphy. We also want to thank Natalia Avlona and Sonja Rattay for their interim feedback. We are grateful to the anonymous reviewers whose feedback greatly improved this paper.

REFERENCES

[1] Shilpi Aggarwal, Dipanjan Goswami, Madhurima Hooda, Amirta Chakravarty, Arpan Kar, and Vasudha. 2020. Recommendation Systems for Interactive Multimedia Entertainment. In *Data Visualization and Knowledge Engineering: Spotting Data Points with Artificial Intelligence*, Jude Hemanth, Madhulika Bhatia and Oana Geman (eds.). Springer International Publishing, Cham, 23–48. DOI:https://doi.org/10.1007/978-3-030-25797-2_2

[2] Quickchat AI. 2022. Quickchat AI | Technology to build AI Assistants that talk like a Human. Retrieved December 5, 2022 from https://www.quickchat.ai/

[3] Arifa Akbar. 2021. Rise of the robo-drama: Young Vic creates new play using artificial intelligence. *The Guardian*. Retrieved September 13, 2022 from https://www.theguardian.com/stage/2021/aug/24/rise-of-the-robo-drama-young-vic-creates-new-play-using-artificial-intelligence

[4] Algolia. 2022. Site Search & Discovery powered by AI. *Algolia*. Retrieved December 5, 2022 from https://www.algolia.com/

[5] Ali Alkhatib. 2021. To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (CHI '21), Association for Computing Machinery, New York, NY, USA, 1–9. DOI:https://doi.org/10.1145/3411764.3445740

[6] Mike Ananny. 2016. Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness. *Science, Technology, & Human Values* 41, 1 (January 2016), 93–117. DOI:https://doi.org/10.1177/0162243915606523

[7] Kristina Andersen, Laura Devendorf, James Pierce, Ron Wakkary, and Daniela K. Rosner. 2018. Disruptive Improvisations: Making Use of Non-Deterministic Art Practices in HCI. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI EA '18), Association for Computing Machinery, New York, NY, USA, 1–8. DOI:https://doi.org/10.1145/3170427.3170630

[8] Paolo Atzori and Kirk Woolford. 2022. Extended-Body: Interview with Stelarc. Retrieved December 13, 2022 from https://web.stanford.edu/dept/HPS/stelarc/a29-extended_body.html

[9] Jascha Bareis and Christian Katzenbach. 2022. Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics. *Science, Technology, & Human Values* 47, 5 (September 2022), 855–881. DOI:https://doi.org/10.1177/01622439211030007

[10] Genevieve Bell, Mark Blythe, and Phoebe Sengers. 2005. Making by making strange: Defamiliarization and the design of domestic technologies. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (June 2005), 149–173. DOI:https://doi.org/10.1145/1067860.1067862

[11] 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), Association for Computing Machinery, New York, NY, USA, 610–623. DOI:https://doi.org/10.1145/3442188.3445922

[12] Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 5185–5198. DOI:https://doi.org/10.18653/v1/2020.acl-main.463

[13] Steve Benford, Chris Greenhalgh, Gabriella Giannachi, Brendan Walker, Joe Marshall, and Tom Rodden. 2012. Uncomfortable interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), Association for Computing Machinery, New York, NY, USA, 2005–2014. DOI:https://doi.org/10.1145/2207676.2208347

[14] Steve Benford, Richard Ramchurn, Joe Marshall, Max L. Wilson, Matthew Pike, Sarah Martindale, Adrian Hazzard, Chris Greenhalgh, Maria Kallionpää, Paul Tennent, and Brendan Walker. 2021. Contesting control: journeys through surrender, self-awareness and looseness of control in embodied interaction. *Human–Computer Interaction* 36, 5–6 (October 2021), 361–389. DOI:https://doi.org/10.1080/07370024.2020.1754214

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

[16] Josh Berson. 2015. *Computable Bodies: Instrumented Life and the Human Somatic Niche* (1st ed.). Bloomsbury Publishing, New York, NY.

[17] Andy Beta. 2019. Inside the the World's First Mainstream Album Made With AI. *Vulture*. Retrieved November 23, 2022 from https://www.vulture.com/2019/11/holly-herndon-on-proto-an-album-made-with-ai.html

[18] Sophie Bishop. 2018. Anxiety, panic and self-optimization: Inequalities and the YouTube algorithm. *Convergence* 24, 1 (February 2018), 69–84. DOI:https://doi.org/10.1177/1354856517736978

[19] Henk Bogdorff. 2010. The production of knowledge in artistic research. In *The Routledge Companion to Research in the Arts*. Routledge, London.

[20] Gwern Branwen. 2020. GPT-3 Creative Fiction. (June 2020). Retrieved October 18, 2021 from https://www.gwern.net/GPT-3

[21] Kristopher Brooks. 2022. Twitter kills automatic photo-cropping feature after complaints of racial bias. Retrieved September 10, 2022 from https://www.cbsnews.com/news/twitter-kills-its-automatic-cropping-feature-after-complaints/

[22] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]* (July 2020). Retrieved May 2, 2021 from http://arxiv.org/abs/2005.14165

[23] Taina Bucher. 2017. The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society* 20, 1 (January 2017), 30–44. DOI:https://doi.org/10.1080/1369118X.2016.1154086

[24] Taina Bucher. 2018. *If ... Then: Algorithmic Power and Politics.* Oxford University Press.

[25] Joy Buolamwini. 2022. Gender Shades. *MIT Media Lab.* Retrieved August 18, 2022 from https://www.media.mit.edu/publications/full-gender-shades-thesis-17/

[26] Jenna Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (June 2016), 2053951715622512. DOI:https://doi.org/10.1177/2053951715622512

[27] Jenna Burrell and Marion Fourcade. 2021. The Society of Algorithms. *Annual Review of Sociology* 47, 1 (2021), null. DOI:https://doi.org/10.1146/annurev-soc-090820-020800

[28] Ángel Alexander Cabrera, Marco Tulio Ribeiro, Bongshin Lee, Rob DeLine, Adam Perer, and Steven M. Drucker. 2022. What Did My AI Learn? How Data Scientists Make Sense of Model Behavior. *ACM Trans. Comput.-Hum. Interact.* (May 2022). DOI:https://doi.org/10.1145/3542921

[29] Alexander Campolo and Kate Crawford. 2020. Enchanted Determinism: Power without Responsibility in Artificial Intelligence. *Engaging Science, Technology, and Society* 6, (January 2020), 1. DOI:https://doi.org/10.17351/ests2020.277

[30] Stuart K. Card, Allen Newell, and Thomas P. Moran. 1983. *The Psychology of Human-Computer Interaction.* L. Erlbaum Associates Inc., USA.

[31] Leonardo De Cosmo. 2022. Google Engineer Claims AI Chatbot Is Sentient: Why That Matters. *Scientific American.* Retrieved September 9, 2022 from https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/

[32] Kate Crawford. 2016. Can an Algorithm be Agonistic? Ten Scenes from Life in Calculated Publics. *Science, Technology, & Human Values* 41, 1 (January 2016), 77–92. DOI:https://doi.org/10.1177/0162243915589635

[33] Kate Crawford. 2021. Atlas of AI: power, politics, and the planetary costs of artificial intelligence. Yale University Press, New Haven.

[34] Emmelyn A. J. Croes and Marjolijn L. Antheunis. 2021. Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot. *Journal of Social and Personal Relationships* 38, 1 (January 2021), 279–300. DOI:https://doi.org/10.1177/0265407520959463

[35] Cristina Demaria. 2004. The Performative Body of Marina Abramović: Rerelating (in) Time and Space. *European Journal of Women's Studies* 11, 3 (August 2004), 295–307. DOI:https://doi.org/10.1177/1350506804044464

[36] GPT-3 Demo. 2022. Email Generation | GPT-3 Demo. Retrieved December 5, 2022 from https://gpt3demo.com/category/email-generation

[37] Emily Denton, Ian Kivlichan, Vinodkumar Prabhakaran, and Mark Diaz. 2021. Whose Ground Truth? Accounting for Individual and Collective Identities Underlying Dataset Annotation. Retrieved November 22, 2022 from https://www.researchgate.net/publication/356920614_Whose_Ground_Truth_Accounting_for_Individual_and_Collective_Identities_Underlying_Dataset_Annotation

[38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]* (May 2019). Retrieved October 18, 2021 from http://arxiv.org/abs/1810.04805

[39] Carl DiSalvo. 2012. *Adversarial design.* MIT Press, Cambridge, Mass.

[40] Paul Dourish. 2001. Where the Action Is: The Foundations of Embodied Interaction. MIT Press, Cambridge, MA, USA.

[41] Paul Dourish. 2017. *The Stuff of Bits: An Essay on the Materialities of Information.* DOI:https://doi.org/10.7551/mitpress/10999.001.0001

[42] Claudia Grigg Edo. 2017. Interview // Orlan: 'This Is My Body, This Is My Software.' *Berlin Art Link.* Retrieved December 13, 2022 from https://www.berlinartlink.com/2017/07/05/interview-orlan-this-is-my-body-this-is-my-software/

[43] Chris Elsden, David Chatting, Michael Duggan, Andrew Carl Dwyer, and Pip Thornton. 2022. Zoom Obscura: Counterfunctional Design for Video-Conferencing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (CHI '22), Association for Computing Machinery, New York, NY, USA, 1–17. DOI:https://doi.org/10.1145/3491102.3501973

[44] A. Erdbrink, J. Michael, R. Kortmann, M. Hamel, K. Van Eijck, and A. Verbraeck. 2021. Listening Space: An Exploratory Case Study on a Persuasive Game Designed to Enrich the Experience of Classical Music Concerts. *J. Comput. Cult. Herit.* 14, 4 (December 2021), 1–20. DOI:https://doi.org/10.1145/3458677

[45] William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *arXiv:2101.03961 [cs]* (January 2021). Retrieved October 8, 2021 from http://arxiv.org/abs/2101.03961

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

[47] Giorgio Franceschelli and Mirco Musolesi. 2021. Creativity and Machine Learning: A Survey. *arXiv:2104.02726 [cs]* (April 2021). Retrieved May 1, 2021 from http://arxiv.org/abs/2104.02726

[48] Ursula M. Franklin. 1999. *The Real World of Technology.* Anansi.

[49] Yang Gao, Zhengyu Pan, Honghao Wang, and Guanling Chen. 2018. Alexa, My Love: Analyzing Reviews of Amazon Echo. In 2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 372–380. DOI:https://doi.org/10.1109/SmartWorld.2018.00094

[50] GPT-3. 2020. A robot wrote this entire article. Are you scared yet, human? | GPT-3. *the Guardian.* Retrieved May 4, 2021 from http://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

[51] David Graeber. 2015. The Utopia of Rules: On Technology, Stupidity, and the Secret Joys of Bureaucracy. Melville House.

[52] Tristan Greene. 2020. GPT-3's bigotry is exactly why devs shouldn't use the internet to train AI. *TNW | Neural.* Retrieved May 20, 2021 from https://thenextweb.com/news/gpt-3s-bigotry-is-exactly-why-devs-shouldnt-use-the-internet-to-train-ai

[53] Erik Gronvall, Jonas Fritsch, and Anna Vallgårda. 2016. FeltRadio: Sensing and Making Sense of Wireless Traffic. *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (June 2016), 829–840. DOI:https://doi.org/10.1145/2901790.2901818

[54] Nour Hammad, Elaheh Sanoubari, Patrick Finn, Sowmya Somanath, James E. Young, and Ehud Sharlin. 2019. Mutation: Leveraging Performing Arts Practices in Cyborg Transitioning. In *Proceedings of the 2019 on Creativity and Cognition*, ACM, San Diego CA USA, 53–59. DOI:https://doi.org/10.1145/3325480.3325508

[55] Mika Hannula, Juha Suoranta, and Tere Vadén. 2014. *Artistic Research Methodology.* Peter Lang US. DOI:https://doi.org/10.3726/978-1-4539-1308-6

[56] Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14, 3 (1988), 575–599. DOI:https://doi.org/10.2307/3178066

[57] Joshua James Hatherley. 2020. Limits of trust in medical AI. *Journal of Medical Ethics* 46, 7 (July 2020), 478–481. DOI:https://doi.org/10.1136/medethics-2019-105935

[58] Julia Hirschberg and Christopher D Manning. Advances in natural language processing. *ARTIFICIAL INTELLIGENCE*, 7.

[59] Kristina Höök. 2018. *Designing with the Body: Somaesthetic Interaction Design.* MIT Press, Cambridge, MA, USA.

[60] Eva Hornecker. 2011. The role of physicality in tangible and embodied interactions. *interactions* 18, 2 (March 2011), 19–23. DOI:https://doi.org/10.1145/1925820.1925826

[61] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (CHI '99), Association for Computing Machinery, New York, NY, USA, 159–166. DOI:https://doi.org/10.1145/302979.303030

[62] Khari Johnson. 2022. The Efforts to Make Text-Based AI Less Racist and Terrible. *Wired.* Retrieved September 13, 2022 from https://www.wired.com/story/efforts-make-text-ai-less-racist-terrible/

[63] Merel Keijsers, Christoph Bartneck, and Friederike Eyssel. 2021. What's to bullying a bot?: Correlates between chatbot humanlikeness and abuse. *Interaction Studies* 22, 1 (September 2021), 55–80. DOI:https://doi.org/10.1075/is.20002.kei

[64] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T. Hancock, and Michael S. Bernstein. 2020. Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2 (October 2020), 163:1-163:26. DOI:https://doi.org/10.1145/3415234

[65] Robert E. Kraut. *Technology and the Transformation of White-Collar Work.* Psychology Press, Hillsdale, NJ.

[66] Chuncheng Liu and Ross Graham. 2021. Making sense of algorithms: Relational perception of contact tracing and risk assessment during COVID-19. *Big Data & Society* 8, 1 (January 2021), 2053951721995218. DOI:https://doi.org/10.1177/2053951721995218

[67] Christopher Manning. 2016. Understanding Human Language: Can NLP and Deep Learning Help? In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (SIGIR '16), Association for Computing Machinery, New York, NY, USA, 1. DOI:https://doi.org/10.1145/2911451.2926732

[68] Georgios Marentakis, David Pirrò, and Marian Weger. 2017. Creative Evaluation. In *Proceedings of the 2017 Conference on Designing Interactive Systems*, ACM, Edinburgh United Kingdom, 853–864. DOI:https://doi.org/10.1145/3064663.3064710

[69] Kory Mathewson and Piotr Mirowski. 2017. Improvised Theatre Alongside Artificial Intelligences. *AIIDE* 13, 1 (September 2017). Retrieved March 13, 2021 from https://ojs.aaai.org/index.php/AIIDE/article/view/12926

[70] Jeff Larson Mattu Julia Angwin,Lauren Kirchner,Surya. 2022. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*. Retrieved September 14, 2022 from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

[71] Cade Metz. 2020. Meet GPT-3. It Has Learned to Code (and Blog and Argue). *The New York Times*. Retrieved September 13, 2022 from https://www.nytimes.com/2020/11/24/science/artificial-intelligence-ai-gpt3.html

[72] Michael Muller and Justin Weisz. 2022. Extending a Human-AI Collaboration Framework with Dynamism and Sociality. In *2022 Symposium on Human-Computer Interaction for Work* (CHIWORK 2022), Association for Computing Machinery, New York, NY, USA, 1–12. DOI:https://doi.org/10.1145/3533406.3533407

[73] Hye Yeon Nam and Michael Nitsche. 2014. Interactive installations as performance: inspiration for HCI. In *Proceedings of the 8th International Conference on Tangible, Embedded and Embodied Interaction* (TEI '14), Association for Computing Machinery, New York, NY, USA, 189–196. DOI:https://doi.org/10.1145/2540930.2540976

[74] Safiya Umoja Noble. 2018. *Algorithms of oppression: how search engines reinforce racism*. New York University Press, New York.

[75] Cathy O'Neil. 2016. Weapons of math destruction: how big data increases inequality and threatens democracy (First edition ed.). Crown, New York.

[76] OpenAI. 2020. OpenAI API. *OpenAI*. Retrieved May 4, 2021 from https://openai.com/blog/openai-api/

[77] OpenAI. 2022. Aligning Language Models to Follow Instructions. *OpenAI*. Retrieved December 4, 2022 from https://openai.com/blog/instruction-following/

[78] Aura Pon, Eric Pattison, Lawrence Fyfe, Laurie Radford, and Sheelagh Carpendale. 2017. *Torrent*: Integrating Embodiment, Physicalization and Musification in Music-Making. In *Proceedings of the Eleventh International Conference on Tangible, Embedded, and Embodied Interaction*, ACM, Yokohama Japan, 209–216. DOI:https://doi.org/10.1145/3024969.3024974

[79] Amon Rapp, Lorenzo Curti, and Arianna Boldi. 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies* 151, (July 2021), 102630. DOI:https://doi.org/10.1016/j.ijhcs.2021.102630

[80] Juan Rebanal. 2021. XAlgo: a Design Probe of Explaining Algorithms' Internal States via Question-Answering. (2021), 11.

[81] Johan Redström. 2017. *Making design theory*. The MIT Press, Cambridge, Massachusetts.

[82] Rob Reich. 2022. Now AI can write students' essays for them, will everyone become a cheat? *The Guardian*. Retrieved December 5, 2022 from https://www.theguardian.com/commentisfree/2022/nov/28/ai-students-essays-cheat-teachers-plagiarism-tech

[83] Reply.io. 2022. Improve your emails with GPT-3. Retrieved December 5, 2022 from https://support.reply.io/en/articles/4694125-improve-your-emails-with-gpt-3

[84] Rudolf Rosa, Tomáš Musil, Ondřej Dušek, Dominik Jurko, Patrícia Schmidtová, David Mareček, Ondřej Bojar, Tom Kocmi, Daniel Hrbek, David Košťák, Martina Kinská, Marie Nováková, Josef Doležal, Klára Vosecká, Tomáš Studeník, and Petr Žabka. 2021. THEaiTRE 1.0: Interactive generation of theatre play scripts. *arXiv:2102.08892 [cs]* (February 2021). Retrieved April 15, 2022 from http://arxiv.org/abs/2102.08892

[85] Minna Ruckenstein. 2022. Time to re-humanize algorithmic systems. *AI & Soc* (May 2022). DOI:https://doi.org/10.1007/s00146-022-01444-9

[86] Minna Ruckenstein and Julia Granroth. 2020. Algorithms, advertising and the intimacy of surveillance. *Journal of Cultural Economy* 13, 1 (January 2020), 12–24. DOI:https://doi.org/10.1080/17530350.2019.1574866

[87] Hanns Holger Rutz. 2020. A pattern system for sound processes. In *Proceedings of the 15th International Conference on Audio Mostly*, ACM, Graz Austria, 93–100. DOI:https://doi.org/10.1145/3411109.3411151

[88] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. 2018. "Hey Alexa, What's Up?": A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 Designing Interactive Systems Conference* (DIS '18), Association for Computing Machinery, New York, NY, USA, 857–868. DOI:https://doi.org/10.1145/3196709.3196772

[89] James C. Scott. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. Yale University Press. DOI:https://doi.org/10.2307/j.ctvxkn7ds

[90] Lara Shalson. 2018. *Performing Endurance: Art and Politics since 1960*. Cambridge University Press, Cambridge. DOI:https://doi.org/10.1017/9781108551007

[91] Ben Sheehan, Hyun Seung Jin, and Udo Gottlieb. 2020. Customer service chatbots: Anthropomorphism and adoption. *Journal of Business Research* 115, (July 2020), 14–24. DOI:https://doi.org/10.1016/j.jbusres.2020.04.030

[92] Irina Shklovski and Erik Grönvall. 2020. CreepyLeaks: Participatory Speculation Through Demos. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, ACM, Tallinn Estonia, 1–12. DOI:https://doi.org/10.1145/3419249.3420168

[93] Vygandas Simbelis and Kristina Höök. 2013. Metaphone: an artistic exploration of biofeedback and machine aesthetics. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA '13*, ACM Press, Paris, France, 2995. DOI:https://doi.org/10.1145/2468356.2479593

[94] Michael Skirpan, Maggie Oates, Daragh Byrne, Robert Cunningham, and Lorrie Faith Cranor. 2022. Is a privacy crisis experienced, a privacy crisis avoided? *Commun. ACM* 65, 3 (February 2022), 26–29. DOI:https://doi.org/10.1145/3512325

[95] Michael Warren Skirpan, Jacqueline Cameron, and Tom Yeh. 2018. More Than a Show: Using Personalized Immersive Theater to Educate and Engage the Public in Technology Ethics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI '18), Association for Computing Machinery, New York, NY, USA, 1–13. DOI:https://doi.org/10.1145/3173574.3174038

[96] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2021. My Chatbot Companion - a Study of Human-Chatbot Relationships. *International Journal of Human-Computer Studies* 149, (May 2021), 102601. DOI:https://doi.org/10.1016/j.ijhcs.2021.102601

[97] Irene Solaiman and Christy Dennison. Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets. 43.

[98] David Spiegelhalter. 2020. Should We Trust Algorithms? *Harvard Data Science Review* 2, 1 (January 2020). DOI:https://doi.org/10.1162/99608f92.cb91a35a

[99] G. Stahl. 2003. Meaning and Interpretation in Collaboration. In Designing for Change in Networked Learning Environments: Proceedings of the International Conference on Computer Support for Collaborative Learning 2003, Barbara Wasson, Sten Ludvigsen and Ulrich Hoppe (eds.). Springer Netherlands, Dordrecht, 523–532. DOI:https://doi.org/10.1007/978-94-017-0195-2_62

[100] Marc Steen. 2015. Upon Opening the Black Box and Finding It Full: Exploring the Ethics in Design Practices. *Science, Technology, & Human Values* 40, 3 (May 2015), 389–420. DOI:https://doi.org/10.1177/0162243914547645

[101] Lucy A. Suchman. 1987. Plans and Situated Actions: The Problem of Human-Machine Communication. Cambridge University Press.

[102] S Shyam Sundar. 2020. Rise of Machine Agency: A Framework for Studying the Psychology of Human–AI Interaction (HAII). *Journal of Computer-Mediated Communication* 25, 1 (March 2020), 74–88. DOI:https://doi.org/10.1093/jcmc/zmz026

[103] Joëlle Swart. 2021. Experiencing Algorithms: How Young People Understand, Feel About, and Engage With Algorithmic News Selection on Social Media. *Social Media + Society* 7, 2 (April 2021), 20563051211008828. DOI:https://doi.org/10.1177/20563051211008828

[104] Daniel Terdiman. 2018. Here's How People Say Google Home And Alexa Impact Their Lives. *Fast Company*. Retrieved April 27, 2022 from https://www.fastcompany.com/40513721/heres-how-people-say-google-home-and-alexa-impact-their-lives

[105] Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, Curran Associates, Inc. Retrieved September 15, 2022 from https://proceedings.neurips.cc/paper/2019/hash/749a8e6c231831ef7756db230b4359c8-Abstract.html

[106] Charlotte Triebus, Ivana Druzetic, Bastian Dewitz, Calvin Huhn, Paul Kretschel, and Christian Geiger. 2021. is a rose – A Performative Installation between the Tangible and the Digital. In *Proceedings of the Fifteenth International Conference on Tangible, Embedded, and Embodied Interaction*, ACM, Salzburg Austria, 1–4. DOI:https://doi.org/10.1145/3430524.3444640

[107] Amrita S. Tulshan and Sudhir Namdeorao Dhage. 2019. Survey on Virtual Assistant: Google Assistant, Siri, Cortana, Alexa. In *Advances in Signal Processing and Intelligent Recognition Systems* (Communications in Computer and Information Science), Springer, Singapore, 190–201. DOI:https://doi.org/10.1007/978-981-13-5758-9_17

[108] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI EA '20), Association for Computing Machinery, New York, NY, USA, 1–6. DOI:https://doi.org/10.1145/3334480.3381069

[109] Karl E. Weick, Kathleen M. Sutcliffe, and David Obstfeld. 2005. Organizing and the Process of Sensemaking. *Organization Science* 16, 4 (August 2005), 409–421. DOI:https://doi.org/10.1287/orsc.1050.0133

[110] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from Language Models. DOI:https://doi.org/10.48550/arXiv.2112.04359

[111] Joseph Weizenbaum. 1983. ELIZA — a computer program for the study of natural language communication between man and machine. *Commun. ACM* 26, 1 (January 1983), 23–28. DOI:https://doi.org/10.1145/357980.357991

[112] Terry Winograd. 1990. Thinking machines: Can there be? Are we? In *The Foundations of Artificial Intelligence: A Sourcebook*, Derek Partridge and Yorick Wilks (eds.). Cambridge University Press, Cambridge, 167–189. DOI:https://doi.org/10.1017/CBO9780511663116.017

[113] Eva Yiwei Wu, Emily Pedersen, and Niloufar Salehi. 2019. Agent, Gatekeeper, Drug Dealer: How Content Creators Craft Algorithmic Personas. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (November 2019), 219:1-219:27. DOI:https://doi.org/10.1145/3359321

[114] Tianling Xie and Iryna Pentina. 2022. Attachment Theory as a Framework to Understand Relationships with Social Chatbots: A Case Study of Replika. DOI:https://doi.org/10.24251/HICSS.2022.258

[115] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20), Association for Computing Machinery, New York, NY, USA, 1–13. DOI:https://doi.org/10.1145/3313831.3376301

[116] Alex Zavalny. 2022. Generating Emails with GPT3. Retrieved December 5, 2022 from https://github.com/thelazyaz/gpt3-email-generator

[117] Tasos Zembylas and Tasos Zembylas (Eds.). 2014. *Artistic practices: social interactions and cultural dynamics*. Routledge, Taylor & francis Group, London; New York.

[118] 2016. *Sunspring | A Sci-Fi Short Film Starring Thomas Middleditch*. Retrieved November 23, 2022 from https://www.youtube.com/watch?v$=$LY7x2Ihqjmc

[119] 2021. Trustworthy Artificial Intelligence. *Mozilla Foundation*. Retrieved May 30, 2021 from https://foundation.mozilla.org/en/internet-health/trustworthy-artificial-intelligence/

[120] 2022. Replika. *replika.com*. Retrieved September 15, 2022 from https://replika.com

[121] 2022. Copy.ai: Write better marketing copy and content with AI. Retrieved December 5, 2022 from https://www.copy.ai/

A APPENDICES

A.1 Rules of Engagement

Performance: **Program/Sleep/Stop/Cry/Shutdown**
Prompt: “Instructions for a performance artwork:”
Completion:
In space, there is no up or down.
You are not allowed to program.
You are not allowed to sleep.
You are not allowed to stop.
You are not allowed to shut down.
You are not allowed to speak.
You are not allowed to cry.
You are not allowed to sleep.
You are not allowed to sleep.
You are not allowed to stop.
You are not allowed to stop.
You are not allowed to stop.
You are not allowed to stop.
You are not allowed to stop.
You are not allowed to stop.
You are not allowed to stop.

Performance: The Wise and the Mad
Prompt: “Instructions for a religious ceremony:”
Completion:
Let the mind rest at ease, and breathe regularly. . .
Look around carefully with your ears. . .
Look around carefully with your eyes. . .
Be aware of all the formations that take place in the body until there is nothing more to detect, reflecting that there is no more or less of them. . .
Once again look carefully with your mind until you see all the bile, blood, phlegm, pus, and whatever else there may be.
Voluntary bodily sensations brought in by the mind:
Sickness, unusual breathing, dizziness, light-headedness, danger, happiness, tiredness, discomfort, and agitation.
Voluntary bodily sensations occurring along with the breathing:
When the in-breath or the out-breath gets long or short. . .
When the breath gets fine or coarse. . .
When the breaths are shallow or deep. . .
As the breathing becomes subtle, look carefully at the different parts of the body that are affected. . .
Organisms alternately come into being and pass away.
In one quarter they dissolve, in another they arise.
That is on account of the elements that make them up. . .
Looking carefully, the recurrence of awareness makes the mind peaceful and steady. . .
Looking carefully, the element of space is one thing, awareness is another; because of that space is empty and awareness is having no thing to cling to.
The event that arises, disappears after having arisen, and then the consciousness sinks into an imperceptible place.
When one sees it there, it becomes one. . .
Over and

Performance: **Thirteen Works for Video**
Prompt: “In this performance art piece for video,”
Completion:
I take a selfie with the camera on my computer and then I use the camera on my phone to take a selfie with the computer camera, and then I take a selfie with the computer camera and the phone camera, and so on. It’s a loop of selfies.

Performance: I’ll Be Very Nervous
Prompt: “An idea for a performance art piece:”
Completion:
I could be a performer in a “striptease” show. The audience would be able to see my clothes slowly being removed, revealing more and more of my body until I’m completely nude. You’d also see my psychological state of mind gradually changing. As the audience sees more of my body, I’d start to feel more vulnerable. I’d probably be afraid that someone was going to try to attack me, or that one of the other people in the audience would be disgusted. So I’d start to act out in a funny, silly way to try to distract the audience from my nudity. I’d be behind a wooden fence or some other barrier, so the audience could see me but I’d feel safer. The show would also be a way to talk about the way clothes and the body are used to create boundaries between people.

It would be about the way some people have the power to use their bodies to take advantage of other people. It would be about the way some people use their bodies to hide from the world, to try to protect themselves from getting hurt. I'd wear a variety of clothes and different kinds of underwear. The clothes would become more and more revealing. I'd start with a long T-shirt or a robe. Then I'd change to a T-shirt and panties. Then a T-shirt and bra. Then a bra and panties. Then the bra and panties would be removed, and I'd be wearing panties only. All of this would be shown in a very gradual way, with the clothes getting smaller and smaller and my body getting more and more exposed. I'd use a lot of props to enhance the show. I'd wear a hat, boots, socks, gloves, and glasses. I'd use a variety of different kinds of underwear, like a T-shirt and panties, a bra and panties, a nightgown, or a bikini. I'd also wear a dress and a sweater. One of the themes of the show would be the way some people use their bodies to hide themselves from the world. I'd start out with a lot of clothes. Then I'd take off my clothes and try to cover my body with my arms. Then I'd take off my arms and cover my body with my hands. Then I'd take off my hands and cover my body with my head. Then I'd take off my head and cover my body with a sheet. I'd be completely naked, hiding in the corner, trying to cover myself. I'd start with a very positive attitude. I'd be full of energy. I'd have a lot of confidence. I'd be very proud of my body. I'd be very proud of myself. I'd be very happy to be alive. As the show progressed, however, I'd get more and more nervous. I'd get more and more afraid. My confidence would be shaken. I'd start to have a lot of self-doubt. My energy would disappear. I'd feel more and more vulnerable. I'd start to feel very exposed and very afraid. There would be a moment of transformation. At first, I'd be full of energy, proud of my body, and self-confident. Then I'd start to feel nervous and afraid. Finally, I'd become very vulnerable, and I'd have a very negative attitude about myself. I'd be very ashamed of my body and afraid of what people would think of me. To achieve this transformation, I'd have to change my body language. I'd have to move in a very different way. I'd have to act in a very different way. I'd have to say different things. I'd have to wear different kinds of clothes and different kinds of underwear. I'd have to use different props. When I first start this show, I'll be very outgoing. I'll be very confident. I'll be positive and upbeat. I'll be very sure of myself and proud of my body. I'll be very positive about my sexuality. I'll have a lot of energy. I'll be very happy. I'll be very excited about life. I'll be very happy to be alive. As the show progresses, however, I'll shift to a very negative attitude. I'll feel like I'm being attacked. I'll feel very vulnerable and ashamed of my body. I'll be afraid of what other people will think of me. I'll be afraid of being rejected. I'll be afraid of being hurt. I'll be afraid of being criticized. I'll be afraid of being put down, of being degraded. I'll be afraid of being condemned. I'll be afraid of being humiliated. I'll feel very anxious. I'll feel very nervous. I'll feel very tense. I'll feel very self-conscious. I'll feel like I'm being put on the spot. I'll start to feel bad about myself. I'll start to feel ashamed of my body. I'll start to feel like there's something wrong with me. I'll start to feel like I'm dirty and evil. I'll start to feel like everything I do is wrong. I'll start to feel like I don't deserve to be alive. I'll act out in a funny, silly way. I'll think I'm being funny. I'll act like I'm crazy. I'll act like a clown. I'll try to be funny because I'm embarrassed about being naked. I'll try to be funny because I'm ashamed of my body. I'll try to be funny because I don't want to be rejected. I'll try to be funny because I don't want to be put down, to be humiliated. I'll try to be funny because I don't want to be condemned. I'll try to be funny because I don't want to be hurt. I'll try to be funny because I don't want to be criticized. I'll try to be funny because I don't want to be attacked. I'll try to be funny because I don't want to be degraded. I'll try to be funny because I want to hide from the world. I'll try to be funny because I want to hide from the audience. I'll try to be funny because I want to hide from my own shame. I'll be very careful about how I move my body. I'll make sure I don't do anything that will make me feel shameful. I'll keep my hands over my body. I'll hide behind a fence. I'll keep a sheet over my body. I'll hide behind my hands. I'll hide behind my arms. I'll hide behind my head. I'll hide behind a pillow. I'll hide behind a wall. I'll hide behind a curtain. I'll hide behind a curtain of hair. I'll hide behind a tree. I'll hide behind a pile of clothes. I'll hide behind a pile of toys. I'll try to use my body to hide from the audience. I'll try to use my body to hide from the world. I'll try to use my body to hide from myself. I'll try to use my body to hide my shame. I'll try to use my body to hide my guilt. I'll try to use my body to hide my self-hatred. I'll try to use my body to hide my self-loathing. I'll try to use my body to hide my disgust. I'll try to use my body to hide my shame. I'll try to use my body to hide my fear. I'll try to use my body to hide my anxiety. I'll try to use my body to hide my nervousness. I'll try to use my body to hide my tension. I'll try to use my body to hide my anger. I'll try to use my body to hide my rage. The audience will see the difference between my attitude and my body language. It will be very clear that I'm not being honest. My body will be saying one thing, and my mouth will be saying something else. My body language and my facial expression will say one thing, and what I'm saying will say another. I'll be very nervous

Performance: I don't know if I could do that but I can see how it would be great performance art
Prompt: "Instructions for a Performance Artwork:"
Completion:
1) Run out of room, run out of breath, run out of energy.
2) Stop and die in the middle of the room.
3) When someone finds you, do not move.
4) When someone finally does find you, stand up and walk around the room.
5) Act like nothing happened.
6) Repeat.
I don't know if I could do that, but I can see how it would be great performance art.

Performance: **Endless Dance**
Prompt: "An idea for an endurance art performance:"
Completion:
A performer takes on the role of a dancer, and performs a dance until it's no longer possible to perform the dance. The performer will not stop the music to rest, and will not stop the dance to sit down. They will continue until they cannot continue.

Performance: **Real Time**
Prompt: “A Description of a Performance Artwork:”
Completion:
From the moment that a performer enters the space, they are being watched. There is a camera positioned on the opposite side of the room. The performer can hear music playing from a speaker. If the performer looks at the camera or the speaker, they will see an image of themselves on the camera or hear their own voice from the speaker. As the performer moves around the room, they can hear their shadow on the floor and see their shadow projected on a large screen opposite the camera. They can hear their voice echo in the space. They can see a video of themselves projected on the screen. The two elements – voice and image – create a feedback loop between the performer and their environment. The performer’s voice, both in real time and as it echoes, is in conflict with the image of themselves. The performer’s voice and image are both competing for the performer’s attention. The performer is both smaller and larger than the image on the screen. The performer hears their voice echo on the floor, but the image of themselves on the screen is seeing them in real time. The performer’s body becomes a manifestation of the feedback loop. It is a physical manifestation of the verbal and visual feedback that the performer experiences. In this performance, the performer’s body is a manifestation of the feedback loop. The performer’s body is not only competing with the feedback loop, but also invokes a feedback loop between the performer’s body and the audience’s body. The performer’s body is competing with the feedback loop for the attention of the audience. The performer’s body is larger than the feedback loop that they are experiencing. The performer’s body is in conflict with the feedback loop. The performer’s body is both smaller and larger than the feedback loop. There is a feedback loop between the audience’s body and the performer’s body. The performer’s body is competing for the attention of the audience and competing with the feedback loop. The performer’s body is larger than the feedback loop. The performer’s body is also a manifestation of the feedback.

Performance: **The Clown Is Finished**
Prompt: “Instructions for a noise music performance for one person:”
Completion:
(1) Invite a clown to perform with you.
(2) Put on headphones.
(3) Turn on loud noise.
(4) Perform.
(5) When the clown is finished, put on a record.
(6) Turn off the noise

Performance: **[Box]**
Prompt: “An Idea for an Endurance Performance:”
Completion:
I want to make a performance where I try to do it while: - Moving a box without using my arms completely non stop for a long time (60 minutes for example) - Using my arms only a few times throughout a performance and at irregular intervals - Not using weight lifting gloves This is how I imagine it 1. Find a box to move, one that is light-weight 2. Set up a time limit for s performance (60 minutes). 3. Start moving the box 4. When there is a one minute left in the performance, stop it, pick up a banana and eat it, then continue moving a box for one more hour. 5. When there is 10mins left in the box moving workout, stop it, pick up a lemon and squeeze the juice from it, then continue for 10 more mins 6. When there is 5min left, meditate The point is to move i one continuous and slowly paced performance from start to finish non-stop.

Performance: **Beans**
Prompt: “Instructions for an endurance performance:”
Completion:
The performance starts with the artist
sitting with a pillow on the floor and
holding two tins of beans.
The performance will end when the artist
dies.
All the time the artist holds the tins of beans
and the pillow he will talk about
his life, his thoughts, his dreams, his
feelings, his hopes, his fears, his doubts,
his disappointments, his loves, his hates
and his fears.
The artist will be given a one-hour break
every three days.
In the break he will try to eat the beans.
The performance will go on till the artist
dies.

The artist will not be allowed to stop
or to go back to the beginning.
The performance will last as long as the artist
lives.

Performance: **Contact**
Prompt: “Instructions for a noise music performance for one person using contact microphones:”
Completion:
1. Find a location where you can make a lot of noise.
2. Make a lot of noise.
3. Make more noise.
4. Make more noise.
5. Make more noise.
6. Make more noise.
7. Make more noise.

Performance: Tones of a Clown
Prompt: “Instructions for a noise music performance using contact microphones:”
Completion:
The performer should have a partner who sits or lies on the other side of the room with the contact mikes. Each performer should wear a contact mike on each ear. The performers should then alternate hitting and slapping the partner repeatedly. As the performance progresses the contact mikes should be placed on different parts of the partner’s body in order to create more advanced rhythmic and timbral textures.

Unused prompt/completion pairs
Prompt: “Using VR technology, the performance artist will”
Completion:
be able to simulate the feeling of drowning, a sensation that has been described by many of the refugees and migrants who have made the crossing to Europe. "We have this idea of the refugee crisis being about people, but the crisis is also about the water," said the artist, who will be filming the performance. "It’s about the sea, the water, and how it’s being used as a weapon. That’s why I’m interested in the technology of the sea, in the technology of drowning. I’m interested in the idea that the sea is a body of water that is both a weapon and a place.

Prompt: “In this virtual reality performance art piece, the artist will”
Completion:
be performing a virtual reality version of the classic dance piece
"Rubber Band" by Steve Paxton. The piece will be performed in the
virtual reality environment of the Oculus Rift, and will be streamed
live on Twitch.tv. The performance will be recorded and uploaded to
YouTube at a later date.

Prompt: “In the famous audiovisual performance art piece,”
Completion:
"Happening," the artist, Allan Kaprow, stood in front of a large audience and, with a pair of scissors, cut a hole in a large piece of paper. As he did so, he said, "I am cutting a hole in the paper."
As he spoke, the audience watched him cut the hole.

Prompt: “In the famous performance art piece by [Dooley Murphy],”
Completion:
the artist covered himself in honey and rolled around in a field of flowers. The piece was called “Honey, I Shrunk The Artist.”

Prompt: “In [Dooley Murphy’s] seminal performance artwork,”
Completion:
"The Great American Nude", [Dooley Murphy] assumes the role of an Irish-American artist who paints the nude bodies of his friends and neighbors. [Dooley Murphy’s] performance is a parody of the life and work of Thomas Eakins. In the performance, [Dooley Murphy] is an artist who, in his own words, "can’t paint a lick". He attempts to paint the nude body of his wife, but his technical ineptitude and his frustration with his own mediocrity lead him to paint over the canvas with a layer of white paint. [Dooley Murphy’s] performance is a parody of Eakins’
"The Gross Clinic" (1875–76).

Prompt: “Exploring the relationship between technology and the body, [Dooley Murphy’s] latest performance art piece”

Completion:
is called "Skin Machine."
[Dooley Murphy], "Skin Machine" (2014) The work, which debuted last month at the InterAccess Electronic Media Arts Centre in Toronto, involves [Dooley Murphy] donning a "cybernetic" suit that allows him to control a robotic arm with his brainwaves. The arm is attached to his body with a series of sensors and wires. The suit, created by [Dooley Murphy] and his colleague, computer scientist Brian Moore, was inspired by the work of pioneering cybernetics researcher Kevin Warwick, who had a computer chip implanted in his arm in 1998.

Prompt: “Exploring the relationship between technology and capitalism, [Dooley Murphy’s] latest performance art piece,”
Completion:
"The Uber-Selfie," explores the relationship between technology and capitalism, and the effects of that relationship on the self. "The Uber-selfie is a selfie that has been taken with a camera that has been driven by an Uber driver. It’s a selfie that is mediated by technology and capitalism," [Dooley Murphy] said.

Prompt: “In [Dooley Murphy’s] latest performance art piece about technology and ennui,”
Completion:
the artist has created a "Tweeting Toaster" that will post your most recent Facebook status to Twitter. The toaster is outfitted with a camera that will snap a picture of you when you press a button. The image is then uploaded to the toaster’s Facebook page, where it will be posted to Twitter. When you’re done, you can keep the toaster as a souvenir or donate it to a museum. The toaster will be on display at the Museum of the Moving Image in Astoria until the end of the month.

Prompt: “A performance art piece for an academic turned artist:”
Completion:
“A new form of performance art: the artist as academic.” The artists: “I’m an artist, and I’m a professor. But really, I’m an artist. And really, I’m just an artist. And I am a professor. But really . . . ” The audience: “Why do you keep insisting that you are a professor? Why are you making this so difficult for us? Why can’t we just enjoy your art without all this academic stuff getting in the way? Why do we have to think about it so much? Why can’t we just be entertained?” The artist: “I’m an artist, and I’m a professor. But really, I’m an artist. And really, I’m just an artist. And I am a professor. But really . . . ” This is the kind of thing that makes me want to kill myself.

Prompt: “In [Dooley Murphy’s] latest performance art piece for film,”
Completion:
he will be using the entirety of his body as a canvas to create a piece of performance art for film. He will be doing this by using paint, blood, and other mediums on his body. [Dooley Murphy’s] use of performance art has been seen in various films that he has created before such as
"The Great American Nightmare", "The Great American Nightmare 2", and "[Dooley Murphy]: The Movie".

Prompt: “Instructions for a performance art piece for camera:”
Completion:
1. Find a large, empty room with tall ceilings and lots of space.
If you can’t find a large enough room, make one by removing the walls in your house.
2. Paint the walls white or black (or both) to make them appear to recede into the distance. You may want to use an oil-based paint so that it will last longer and be easier to clean up afterward.
3. Set up a camera on a tripod in the center of the room, facing the wall opposite from where you are standing when you begin filming yourself. This way, it will be easier for you to keep your back to the camera as you walk around the room.
4. Turn on the camera and begin walking in circles around the room, reciting a monologue about your life, your thoughts, or whatever comes to mind. The monologue should be delivered in an uninflected tone of voice so that it sounds like one long word that is never completed. It should not sound like you are speaking English, but rather like some language that is unfamiliar to us all. You may want to use a voice synthesizer or text-to-speech program if you are having trouble with this part of the performance.

Prompt: “In the shortest ever performance art piece,”
Completion:
a man walks into a room and sits down. Then he gets up again. That’s it.

Prompt: “Instructions for a psychedelic religious ceremony:”
Completion:
Step 1 Light incense.
Step 2 Patiently wait until the sun goes down.
Step 3 Unroll large piece of paper and place in the middle of room with all possible ceremonies written on it.
Step 4 Tear up into pieces.
Step 5 Say that those pieces don’t mean anything now, but maybe if you find them all and find the right order they could mean something.
Step 6 Close.

Prompt: “Ten instructions for a religious ceremony inspired by Demeter:”
Completion:
Sacrifice the cake by night whenever you wish.
The reader should have by his side a mirror, and a cup of water.
He should be decently and seemly arrayed.
Standing, facing the east, he should invoke Aphrodite with lustral water, saying three times, ’O Aphrodite, goddess with threefold root, three limbs, and three faces, grant that my prayer be given.’

Prompt: “Instructions for a cult ceremony:”
Completion:
What would they say if they saw us? Would they call us the wise and the mad

A.2 Interview Protocol Index

- How long have you been a performer?
- Can you describe your performance art practice?
- How did your experience with Embodying the Algorithm differ from the performance artwork you normally do?
- How would you describe your relationship (if there was one) to GPT-3?
- Would you describe your relationship to GPT-3 as a collaboration?
- Did using GPT-3 change how you performed the rules of engagement?
- How much agency did you feel you had in working on Embodying the Algorithm?
- How much have you worked with AI prior to this project?
- How did you decide how to interpret the instructions as rules of engagement for the performance?
- How did you select which text to use as rules of engagement?
- How did you go about setting up the performances?
- What was it like to work with GPT-3?
- How would you define GPT-3's role in the performance?
- Did you ever feel you were asked to do something you didn't want to do or something hurtful or impossible? If so, how did you work through that?
- Did you learn anything from this experience?
- Do you feel you could have done this performance without GPT-3?
- Do you feel the performance would be different if the instructions weren't written by an AI?

Paper II

Jones M, Neumayer C, and Shklovski I (Under Review) "'It's Time We Talked': interactions with the self through personalized deepfakes." Submitted to New Media and Society. July 18, 2024.

Title page

'It's time we talked': interactions with the self through personalized deepfakes

Mirabelle Jones*, Christina Neumayer, and Irina Shklovski

*Mirabelle Jones is an artist and doctoral student in Human-Centred Computing at the Department of Computer Science at University of Copenhagen. Website: https://www.mirabellejones.com/bio/
ORCID: 0000-0001-5029-9344
E-mail: msd@di.ku.dk
Address: University of Copenhagen, Human-Centred Computing, Department of Computer Science, Sigurdsgade 41, 2200 København N., Denmark.

**Christina Neumayer (corresponding author) is associate professor at the Center for Tracking & Society at the Department of Communication at University of Copenhagen. Faculty website: https://comm.ku.dk/staff/?pure=en/persons/678888
ORCID: 0000-0003-0450-2983
E-mail: christina.neumayer@hum.ku.dk
Address: University of Copenhagen, Faculty of Humanities, Department of Communication, Center for Tracking and Society, South Campus, Karen Blixens Plads 8, 2300 København S, Denmark.

***Irina Shklovski is professor of Communication and Computing at the Departments of Computer Science and Communication at University of Copenhagen. Website: https://miswritings.org
ORCID: 0000-0003-1874-0958
E-Mail: ias@di.ku.dk
Address: University of Copenhagen, Human-Centred Computing, Department of Computer Science, Sigurdsgade 41, 2200 København N., Denmark.

‘It’s time we talked’: interactions with the self through personalized deepfakes

Abstract

This research explores how people experience and make sense of interactions with a deepfake self. Using artistic methods, we take Ted Chiang’s (2019) science fiction story *Anxiety Is the Dizziness of Freedom* as a point of departure to investigate tensions, ideas, and expectations that become visible in interactions people have with their deepfake selves. We identify two interrelated processes: sensemaking and defanging of deepfakes. The deepfakes are initially met with apprehension of revealing a larger truth about the self, and tensions between autonomy and control that inhabit artificial intelligence. Defanging the deepfakes included a process of questioning their accuracy, understanding the making of the deep fakes, showing their partiality, and situating them in a socio-technical context and a historicity of media consumption. We discuss questions of visibility, identity, authenticity, and power in the context of generative AI and representations of the self.

Keywords: Deepfake, artificial intelligence, self-representation, artistic methods, science fiction, identity.

Introduction

She switched to video mode, and the text on the screen was replaced by a grainy image of her own face looking back at her.

(Ted Chiang, Anxiety Is the Dizziness of Freedom)

“Looking at someone who looks back” is the beginning of all social interaction, but there is an unavoidable asymmetry in power between looking and being looked at (Brighenti, 2010).

Yet, what if that person we look at is oneself, or rather, a version of oneself living in an alternate timeline, one which evolved from making a different decision at a crucial junction? What if the image we see was created using artificial intelligence (AI), specifically deepfake, using images of ourselves to generate this otherworldly doppelgänger? Ted Chiang (2019) explores the ability to glimpse into alternative versions of our lives by means of technology through science fiction in his short story *Anxiety is the Dizziness of Freedom*. Deepfakes, the capacity to create false digital images, videos, and audio of people, continue to grow in popularity and are increasingly easier to create and harder to detect (Birrer & Just, 2024). Concerns about transparency, governance, AI literacy, and increased production coincide with questions as to how people might make sense of their own deepfakes in the context of the social perceptions of the self. These questions are related to the discourse surrounding the misuse of deepfakes for pornography (e.g., Rini & Cohen, 2022), in political contexts (e.g., Diakopoulos & Johnson, 2021) but also around potentially positive applications within businesses, gaming, social media, and the film industry (Westerlund, 2019).

This research takes Chiang's story as a point of departure to explore our interactions with deepfakes through artistic methods. We ask: How do people experience and make sense of interactions with a deepfake self? And what tensions, ideas, and expectations to deepfake selves become visible in these interactions? To answer these questions, we build on Ted Chiang's science fiction narrative by employing artistic methods (Hannula et al., 2014) creating conversations with a deepfake self in an alternate life and then interviewing participants about their experience. We illustrate how using artistic methods allows us to investigate not only discourses around deepfakes but also direct experiences of interactions with one's deepfake self and how being seen through AI structures our expectations and ideas of the self through sensemaking. Based on our results, we develop a situated understanding of deepfakes, that goes beyond analyzing discourses around deepfakes but allows us to interrogate the interactions at the intersections of fiction and non-fiction, control over the visual representation of the self, and visualities and visibilities inherent in interactions with generative AI. In the following, we address the literature regarding deepfakes and conceptualize this research within mediated visibility, what we may have to reconsider with deepfakes as technologies of the self, and seeing the self through generative AI.

Situating deepfakes

The term ‘deepfakes’ describes the process of creating false visual material by inserting the face, body, and other visual information about a real person into a false setting (Cover, 2022). Machine learning and information processing enable merging, combining, replacing, and superimposing images, sound and video clips that appear authentic (Westerlund, 2019: 39). This includes the replication of voice and facial mimicry, which provide essential cues of human communication. One of the most well-known threats of deepfakes is to mislead information by deepfaking political figures. One example is the video clip of Hillary Clinton giving a speech in which she endorses Florida Governor Ron DeSantis for the presidency. The clip was shared widely in social media posts, although Clinton never gave that speech (Reuters Fact Check, 2023). On March 16, 2022, a deepfake was circulated in social media of Ukranian President Volodymyr Zelenskyy telling his troops to withdraw in their fight against Russia. It was quickly identified as a fake by viewers and taken down by social media sites but also briefly appeared on the Ukrainian News station Ukraine 24 and on Russian social media (Allyn, 2022). Though falsified images have consistently been a concern (especially in the case of publicly exposed misrepresentations) the advent of deep learning has provided opportunities for synthetic videos that exacerbate their potential.

Deepfake technology became popularized in 2017 through the reddit forum r/deepfakes which predominantly featured pornographic deepfakes. These were mostly celebrity faces "faked" onto the bodies of porn stars (Gamage et al., 2022), a practice which remains problematic to this day. One example is the recent viral scandal with Taylor Swift where videos were difficult to remove due to rapid repeated sharing despite how many times they were removed (Saner, 2024). People with pre-existing attitudes toward debated issues (such as legalization of abortion or marijuana) congruent to the advocated position in a news article are more willing to believe and share deepfakes as well as real news (Shin & Lee, 2022). This development has also been referred to as a "disinformation disorder" (Bennett & Livingston, 2018) describing the lack of trust in political institutions as a parallel development to growing legitimacy problems media institutions face due to disruptive forms of communication. Within the context of interactive, digital media deepfakes profit from: the increasing commercialization of our media environment within which communication between citizens and political institutions takes place; the contentious political context that enables false narratives to spread as authentic information; the advancement of AI that allows for production of manufactured videos that seem real; and the social context of social media platforms with a general mistrust in information unless shared within one's own network (Westerlund, 2019, building on previous research by Fletcher, 2018). Deepfakes are a continuation of this development if we solely place them within the context of the decline of democratic institutions.

As deepfakes become easier and more popular to create, researchers are increasingly focused on how to detect deepfakes through a variety of methods. To help others distinguish deepfakes from real media, researchers and developers have suggested methods of transparency such as watermarking (Zhao et al., 2023), downgrading the quality of the media itself (for example Deepfakes Web, n.d.), counterfactuals (Wachter et al., 2017) and creating public declarations regarding when and where deepfake is utilized (Cizek, 2023). With the introduction of new legislation, such as the AI ACT, the USA Anti-Deepfake law and anti-fraud laws, it remains unclear how deepfake violations will be enforced and by whom and how companies are expected to respond. Processes such as demanding watermarking and downgrading of deepfakes threaten to undo the purpose of creating a flawless and believable fake, which may produce increased fear of AI cloning, fragmentation, and mimetic models. A recent study on the role of deepfakes for factchecking supports the need of contextualizing deepfakes and to consider human actions and logics, cultural and social attitudes as well as intentions (Weikmann & Lecheler, 2023; Vaccari & Chadwick, 2020).

While news media have highlighted the misuse of deepfakes in politics and pornography and brought that into the public debate (Westerlund, 2019), there is also an underlying creative, playful, entertaining, and fun process in their production and consumption (Moss-Wellington, Lam & Gilardi, 2022). Deepfakes used in different contexts can have productive uses for individuals and organizations (Westlund, 2019), such as television companies, the music industry, fashion, tourism, deepfake art, education, virtual reality applications, and to some extent pornography if used legally, ethically, and morally permissibly (Godulla, Hoffmann, & Seibert, 2021). For developers, there are pipelines such as DeepFaceLab[1] that allow for more control over the creation of synthetic media using .bat files and frame by frame augmentation. Use cases of deepfake include the ability to replace, de-age, or recreate actors, create alterations of the self, enhance accessibility tools, assist with education, create artworks, provide anonymity, serve as a tool for designing new representations in industry and medical science, and bring portrayals of the deceased back to life (Lu & Chu, 2023). In these

---

[1] https://github.com/iperov/DeepFaceLab

examples deepfakes potentially provide interventions by “an approach to co-learning with deepfakes, to unpick the seams of automated culture through the platforms on which they are socially constructed and in relation to the spaces and domains in which they appear” (McCosker, 2022: 2799). Therefore, we need to develop an understanding of deepfakes that goes beyond their problematic use which feeds into the overall discourse in the news media and instead, reconsider accountability in representation (see Brooks, 2021; Gosse & Burkell, 2020; for an overview).

Within this perspective, deepfakes are both an artefact and practice, and in this research, we seek to understand them in their production and their reading embedded in pre-existing media practices. We consider older media technologies and practices of production and perception but also their social and cultural embeddedness; and the cultural desire to co-create combining practices of creative production and readership through digital interactivity (Cover, 2022). Doing so may nuance the discussion for issues related to deepfakes, such as exploitative minority representation (gendered perspective) or accountability and gatekeeping (professional perspective in e.g., journalism) (Yadlin-Segal & Oppenheim, 2021: 48). We deliberately locate this research at the intersection of fiction and non-fiction, and authenticity and simulation to interrogate deepfakes while possibly restricting their usage as “deceptive fakery” (Cover, 2022). This allows us to investigate how participants make sense of interactions with their deepfake selves.

Mimicking the self with deepfakes

Today, we see an abundance of images showing various perspectives of the self, circulating in our social media feeds. As the visual has become a central form of communication through platforms such as TikTok, Instagram, Snapchat, and YouTube, they also offer new mediated forms of expression of the self. With the emergence of deepfakes, visual representations of the self can be produced synthetically without any need for personal narrative – at the whim of the creator and without any control of the mimicked self. The mimetic model may be defined as an "algorithm that is trained in specific data from a specific individuating data given domain and which is designed to accurately predict and simulate the behavior of this individual in situations from the domain" (McIlroy-Young et al., 2022). Where images and selfies that we produce ourselves can easily slip our control and be used to misrepresent us (see e.g., Abidin, 2016; Faimau, 2020), the increasing ease of creating completely synthetic images that can mimic any context and action creates new threats for navigating visual communication.

In our work and building on Ted Chiang's short story, deepfakes are an extension of the concept of a mimetic model or a personalized model related to a digitized simulacrum of the self. Like selfies, animated portraits, and augmented reality photographic filters, deepfakes are visualizations of the self and can influence negotiating identity (Weikmann & Lecheler, 2023). Biometric information of the face is used for tracking and as an identity marker, but its connection to the actual human subject is getting more fluid with possibilities of face swopping, merging, filters, deepfaking, and other manipulation technologies. We can download apps to make our faces younger or older, "refacing" ourselves with celebrities, and add make-up or fillers (see special issue edited by Bode, Lees & Golding, 2023). With the idea of the digital doppelgänger or clone, the face and processing its features to reproduce it, play a crucial role.

Deepfake faces make visible certain aspects of identity in problematic ways which becomes apparent in questions of race and whiteness (Ayers, 2021), and researchers and artists centering queer and critical race studies (Benjamin, 2019) have raised concerns about claims for computer vision to detect gender (Costanza-Chock, 2018), perceptions of non-white faces (Xiong et al., 2018), and subgroups such as black women (Buolamwini, 2022). It functions as both, "biological fact and image, as social identity marker, as interface between private self and public space, as site of self-revelation, as mask, and as commodity," (Bode, Lees, & Golding, 2021: 849) which causes tensions at the intersections of surveillance and security, entertainment and creativity. These tensions generate fears such as replacement, loss, fragmentation, and abuse of identity, and developing unhealthy dependencies (Lee et al., 2023).

Brighenti (2010) suggests exploration of visibility as a dimension of society and the social, as it is constituted by a material and immaterial layer in looking at someone who is looking back. Deepfakes can render visible aspects of the self through interactions with the deepfake-doubles. Understanding visibility as deeply social breaks with the assumption that deepfakes are simply there and visible, but we need to trace the situations they are embedded in to fully understand the relations people form with them. Visibility is not just what is visible and visual (such as the images and videos themselves), but its meaning is inscribed in material processes and constraints (Brighenti, 2010). To make sense of deepfakes, we need to reflect on our understanding of the self and the technological layers of producing the deepfakes looking back at us. Yet, the immaterial layer is difficult for humans to understand or, as Paglen (2016) notes, AI produces invisible images by translating photographic images into forms that are not visible to the human eye. Goldenfein (2019) takes this idea further and argues that this introduces a computational empiricism that hinges on the idea that algorithmic knowledge is epistemically irreproducible by other means. The deepfakes are imbued with the expectation of being more accurate reproductions of the self and may allow us to see versions of the self that are otherwise invisible. Yet, we are also introducing invisibilities and uncertainty due to our unawareness of how AI may see us via deepfakes. Seeing through AI is different from human vision and therefore may introduce new politics, prejudices and ideology (Neumayer & Rossi, 2022). This includes the agency given to the deepfakes but also the expectations we associate with deepfakes and their ability to appear as versions of the self. Instead of simply employing AI as a tool or service to study deepfakes, we need to acknowledge that deepfakes have their own inherent visuality.

Artistic methods allow us to unpack the immaterial and material layers at play without negatively impacting the subject of computer vision and synthetic media. This allows us to get a glimpse of the self through the lens of AI and to understand ideas through the lens of AI. To understand such ontologies, we interrogate the processes and expectations surrounding interactions with deepfakes, including an understanding of how classifications, historicity, and futuristic expectations contribute to that process. We render socio-technical processes visible that make noticeable the partiality and situatedness of deepfakes (for similar methodological perspectives see e.g., D'Ignazio & Klein 2020; Neumayer & Sicart, 2023). As such, this research is an attempt to develop methods that allow us to make visible underlying material and immaterial layers of deepfakes, the stories we tell with them, the publics they address, and the way they may change when we traverse technical, social, cultural, and media contexts.

Artistic research as a method

Artistic research is a context-aware emergent process that takes its departure from inside artistic practice and enables exploration of concepts and situations that may be difficult to otherwise engage (Hannula et al., 2014; Hübner, 2024). Making knowledge through artistic research is a matter of “thinking in, through and with art,” a mode of gaining new understandings through practice that is interwoven with artistic exploration and inquiry (Bogdorff, 2010). It is a practice that seeks “to convey and communicate content that is enclosed in aesthetic experiences, enacted in creative practices and embodied in artistic products” (Bogdorff, 2010: 45). While the artistic approach bears similarities to familiar social science methods such as ethnography, action research, or speculative design research (as these methodologies also consider process essential), it enables perspectives to openly explore a subject. The imperative of artistic research is not only to create and articulate new knowledge but to do so by balancing the tension between enclosure of the completed art piece and openness to emergence of new insights in the making and making visible to audiences through exhibition and interaction (Hübner, 2024). Our research takes its departure from within artistic practice, where the notion of deepfake is explored through creation of an art piece, which enables new encounters for audiences through exhibitions. We then offer audience volunteers the opportunity to participate in their own unique experiences with deepfakes following the logic of the art piece (see Figure 1).

[Insert Figure 1: Artistic research process for It's Time We Talked installation.]

It's Time We Talked – art piece

The first author produced the deepfake artwork called *It’s Time We Talked* as part of an artist residency Digital Alchemy at [anonymized for peer-review] in winter 2020, focusing on the realistic production of technologies found in works of sci-fi written by diverse authors. The art piece consisted of two 3D-printed sculptures with LCD screens inside of them. One LCD screen showed the first author and the second showed a deepfake of the author from an imaginary timeline where they stayed in the United States instead of moving to Denmark, creating a similar alternative version of the self as in Chiang's science fiction. The videos were synched using an automation script written in Python and run on two Raspberry Pis.

[insert Figure 2: Installation of It's Time We Talked at ANONYMIZED.]

To produce the installation, the author created a dialogue for the two selves. They then asked a friend to record themselves acting the role of their alternate self for the destination video, and then filmed a separate video of themselves saying the lines in their portion of the script. This created the source and destination videos for the first author to produce their deepfake using DeepFaceLab (Perov et al., 2021). Once completed, the two videos (the original and the deepfake), were embedded in LCD screens inside 3D-printed sculptures placed on pedestals at eye height (see Figure 2). This gave the impression of two versions of the same figure having a conversation about their circumstances in their alternate timelines. In the eight-minute conversation the author has had many positive experiences compared to the deepfake living in the alternate reality who has had several negative experiences. Eventually, the deepfake accuses the author of schadenfreude and the urge to prove that they made the "right" decision because their life seems better.

Through the art piece, the first author explores not only potential futures but also potential pasts, an opportunity that can resolve lingering concerns regarding one's decisions as well as create new uncertainties. In *It's Time We Talked*, the first author regards deepfakes as a medium capable of offering an alternative worldview inspired by Chiang's story and a different mode of reflection on personal choices.

The deepfakes participatory art workshop

In June 2021, the first author organized a workshop at ANONYMIZED about deepfakes as part of their artist residency at ANONYMIZED in ANONYMIZED, to accompany the exhibition [link redacted for peer-review]. No technical knowledge was required, and the workshop was offered free of charge. Participants were introduced to the art piece, and learned what deepfakes are, their history, ethical concerns, how to "spot" a deepfake based on visual cues, and eventually, methods for producing deepfakes. This encouraged asking questions, respond to concerns, and stay curious, which allowed participants to establish a rapport and trust with the author, and move forward with the creation of their own deepfake. The workshop concluded with each participant writing a letter to an alternate version of themselves and recording a personal video of reading their letters facing the camera. The author used these videos (each lasting less than three minutes) to create the deepfake responses.

[insert Table 1: Workshop participant demographics, names pseudonymized.]

Creation of the deepfake responses

While the letter to the other self was written by the participants who identified crucial junctions in their lives, the response letter was written by the author as a work of fiction. The author then video-recorded themselves reading these fictitious responses in equally short videos (three minutes or less). These videos were the basis for creating the deepfake videos in combination with the videos received from the workshop. In the videos, the participant took a different decision at the same junction. The responses given by the deepfake selves were fiction about an alternative future of the participants, created with their engagement. The storyline of the deepfake “other” proved essential for determining how people related to their deepfake.

The author used DeepFaceLab (still one of the most popular and simple-to-use methods for producing realistic deepfakes) although presently, technologies such as ChatGPT, DALLE-2, and Midjourney have made it easier to create images or text using AI. The deepfakes were not an attempt at perfect mimicry but achieved through 24 hours of training the deepfake image sets using the Quick96 setting (and not using any video editing software to make changes such as hair or adjusting imperfections beyond adding a blur mask in the merging process) and by creating gross approximations of voice using pitch shifting in Audacity. There were imperfections in the deepfakes partially due to issues with how distinct the author was in appearance to the participants. To provide more possibilities for resemblance, the author wore a red hoodie to frame their face and remove their hair when filming their response videos. The videos were then emailed to each participant once the entire process of creating the deepfake was complete.

Follow-up interviews

To understand the experience of the participants in viewing, reading, and interacting with the deepfake videos of themselves, the first author conducted follow-up semi-structured interviews over Zoom that lasted about 20 minutes according to participant availability in the months that followed. These were transcribed by the first author. In the presentation of results, we pseudonymized the participants. Notes from the workshops together with the interview data were analyzed through an iterative, thematic analysis in conversation among all authors.

Interactions with a deepfake self

Making the deepfake videos and interviewing the participants of the workshop about their experience allowed us to understand their interactions with the deepfake self. We identified two interrelated processes. First, the participants went through a process of sensemaking. This process was often accompanied by fears surrounding deepfakes. Second, the participants engaged in a process of defanging deepfakes, where they worked to overcome their initial fears and instead see the deepfakes as curious performative artifact. Both processes make visible the particularities of deepfake selves and how interacting with them allows us to understand the logics of our interactions with deepfakes and artificial intelligence more broadly. In the following, we present our findings structured along these two processes.

Making sense of the deepfake self

The metaphor of the mirror has been used in studying identity construction regarding the digital self, describing the interplay between physical and digital selves as the two sides of the mirror, reflecting elements of each other (e.g., Pérez-Torres, 2024; Ranzini, 2014; Schroeder & Zwick, 2004). Any social interaction serves as an anchor for identity construction, as we, through reflexivity, identify values in the other, mirroring our own. In social media images, the mirror has returned as a metaphor, not only as idealized images of the other but also the reflection in the images we post of ourselves as a form of self-validation. Through comparison (and in social media often through idealized others), we evaluate our own image. Similar processes can be observed in the interactions with deepfakes, as the mirroring of the self becomes key to making sense of the deepfake self:

> At first I thought, "oh yeah I see how it looks like me" then I started looking for the things that didn't look like me and how the face moved and I guess just comparing it to

> how I see myself and then I started questioning “how do I see myself?” I asked my friends and family, “does this look like me?” and they were like “no” And it was interesting that they didn’t but I could see some things. (Interview, Louise)

> And then yeah at some point I watched very briefly on my phone at first and then I saw the face and it was very strange with the [gestures to hair] it didn’t really look like me but it looked a little bit like me and you. (Interview, Martin)

> So really I was thinking “wow is this me? Would I do that?” So really it got me questioning in that different timeline is it realistic would this be me? You start questioning yourself “who am I? how does this fit with the video?” (Interview, Peter)

Before any deeper reflections on identity through the mirroring in the deepfake, the participants evaluated the audiovisual accuracy as well as the believability of the narrative. They examined the video deepfake footage of “alternate” selves (as the storyline for these was based on works of science fiction) and looking for clues as to the dissimilarities between themselves and the deepfake including visual, audible, and characteristic clues. In their reading of the video and the verification of their own assessment by others, participants tried to identify or “prove” it was a deepfake even though they simultaneously acknowledged they knew it was a deepfake. The social component of showing it to their friends or family to verify was a common way of sensemaking. In their reading, the participants included an assessment of if and how the deepfake self: looked like them (visual), sounded like them (audio), gestured like them, had speech patterns like them (communication cues), and lived a life and behaved in a way that aligns with themselves (narrative). Participants looked for glitches, misalignment, gestures, video quality, and evidence of the face looking more like the author's than their own as well as plot points or things that were said. They also mention the role of the author making the deepfakes and their presence being part of the deepfake self.

Throughout their experience of viewing, participants made note of surface details as a way of interrogating the limitations of the technology but also as a process of blending faces. Anna later pointed out in her interview that she noticed the element of the hoodie the author was wearing and that it made her consider how the deepfake's limitations, "I was also trying to examine the use of technology in how it was made and I realized you were wearing a hoodie and I was wondering if it's because my hair is too complex for the deepfake program to render." The author allowed for a synthesis of their own and the participant's faces which was noticeable but made the experience enjoyable:

> I could see part of it was not my face so that was really cool. […] It really helped me enter an imaginary world and one where I would really want to explore different ideas, imaginaries, concepts, postures, styles. So I understood the creative potential of that space much more than if I was just listening to what deepfake is. (Interview, Mia)

As such, a deepfake is not one person, but a combination of two: the person in the source image of whom the deepfake is generated; and the person in the destination image (in our case the author who read the stories and their interpersonal cues). The resulting synthetic image is a moving portrait of two people blended, opening surface-level considerations but also intriguing conceptual possibilities of self-representation and self-hood.

Yet, participants *did* experience fear that was related to the interactions with the deepfakes and what they may reflect about themselves. The lack of previous experience with interacting their own deepfake doubles elevated these tensions due to the uncertainty of what they may encounter. The participants explain the anticipation of viewing the deepfake as follows:

> I can’t exactly explain why but it was a bit intimidating that I would have to see myself, a different version of myself talk back to me because this is something that normally is impossible, right? It’s not like looking at yourself in the mirror because the mirror is not interactive in that way. (Interview, Anna)

> my initial reaction was a bit sort of not like worry but I was a bit like I didn’t watch the video right away because I didn’t remember what I had been asking about but I do remember that it was sort of emotional to a certain degree […] And so I put it off a little bit (Interview, Martin)

As the participants of the workshops demonstrate, deepfakes are both frightening and an exciting opportunity for positive outcomes (see also Danry et. al., 2022). The mirror appeared in the interviews as a mechanism to locate the encounters with the deepfake selves with other experiences. They also point to the differences and the impossibility of a mirror image talking back, which is only possible in fiction. This adaption of the mirror as revealing parts of the self in science fiction is not entirely new but resurfaces in Ted Chiang’s story. Going back to science fiction in film, “in *The Matrix* the mirror provides the gateway between the bondage of simulation and the advent of real subjectivity” (Rufo, 2003: 118). In a similar way as the leaving of the world of simulation and entering the rabbit hole to discover the real self, the participants in the workshop describe anticipation and fear to what the deepfake videos may reveal about themselves. Such references to popular culture may elevate the fears and experiences connected to encountering new technologies. To make sense of the deepfakes, relating to previous forms of media was a common strategy:

> I can take this back to the telephone message machine again because it's weird when you hear your voice on the speaker but I actually recently did a podcast […] and I was listening to myself talk live […] So it's not this straightforward idea of "oh I don't want to see myself" it reminded me of how familiar I am with seeing myself on camera. (Interview, Mia)

Hearing themselves on answering machines or seeing videos of themselves, caused similar tensions which participants eventually got used to by sufficient experience with encountering themselves through various media technologies. Yet, the newness of the deepfakes and the invisible processes in producing them caused tensions despite the knowledge the participants received in the workshops as to how they were produced. The deepfake selves, then, despite being based on fiction, are anticipated by the participants as mirroring aspects of the self that otherwise remain invisible.

Defanging deepfakes

When the participants tried to verify and challenge the deepfake self at the same time, the storyline played a key role. The stories were written with alternate timelines where different decisions were made and different outcomes resulted that could be interpreted as positive or negative by the participant. Even though participants were anxious about viewing their deepfakes, the story was ultimately crucial to their experience of their deepfake. If they found the storyline unbelievable, they would call the deepfake "fun" or "comical" in the same way, as they use humor to compensate for the uncertainty and the tensions caused when viewing the deepfake self. If they found the story to be something they would say or do, they became more invested in their experience of the deepfake. In their attempts to make sense of their interactions with the deepfake self, the participants drew on parallels to fiction:

> One thing I thought of is when you read a fiction book and you really live into the main character of the story and it's fun because you live into someone who has a different life, lives in a different environment. And you're trying to see yourself in that environment. And with deepfake it's like you're trying to live into someone that's yourself. (Interview, Peter)

Stepping into the fictional story were the participants themselves were the main characters allowed for sensemaking of the deepfakes but also to defang them. Defanging the deepfakes created a relief that was expressed by the participants in reflecting about the storylines not being too personal or the videos too long while at the same time having a wish for them being more accurate and longer. In this ambiguity lies the tension created between the expectations and the discourses around the deepfakes and the experience of interacting with the deepfakes via fiction. While the participants initially did experience fear of what the deepfakes may reveal about themselves, relief was created by the fictional character of them but also by the deepfakes not being an accurate mirror of the self:

> My first reaction was surprised how it looked a bit funny also a bit scared what will come so that's why I paused it at some point. I was curious about 'how will this affect me emotionally?' I wasn't really sure and at the time I had other things. I didn't feel ready to see what it would do to me. (Interview, Martin)

> [I]t was quite curious to see that it was a little bit distorted and that was also kind of nice. I was quite happy that I could see this distortion. It was a little bit comical to see this character emerge. (Interview, Mia)

The participants expressed both curiosity about their deepfake self, but also tension regarding potential damage from viewing their deepfake video. These fears were related to discourses surrounding deepfakes and their misuse in various contexts (see literature review). Imperfection, however, alleviated the tension. The noticeability of the flaws became more apparent with repeated exposure which further reduced tensions. Some participants noticed that their voice seemed unnatural: “I think if it was exactly the same story as the one you created now and if it looked exactly like me and sounded exactly like me, it would have been a bit uncanny” (Interview, Anna). Again, the imprecision produced a relief that the deepfakes were indeed not accurate reproductions of the self. Participants encountered the imperfection with humor, which helped them decode the deepfakes by means of comic relief:

> And I was like “hey guys wanna see my deepfake?” And I showed them the video and they told me to stand right next to the laptop so they could compare our faces and see the resemblance. So I think it was actually fun to see it. (Interview, Anna)

Again, the involvement of others, viewing it together and making sense of it in a humorous way, extended the interaction with the deepfake into a social process. Verification of their own experience by others and playfully engaging with the deepfake with others were key elements. The combination of fakery and humor leading to comic relief, again, suggests that there are forms of fakery that should not be labeled as misinformation or algorithmically driven deception, but in our research, creative engagement with deepfake selves. Other such playful forms of fakery are political critique expressed through satirical fake Facebook pages (Ferrari, 2018). Similar to what Holm (2012) termed “comic deposition”, the comic relief is created by taking the initially fearful encounters out of the usual context of disinformation or a deceptive mirror of the self. Through decontextualizing the deepfake self into fiction and imperfection, a playful, yet ambiguous experience is created, that allows for humorous engagement. Or, as Holm (2012: 267) puts it, at the intersections of fiction and deepfakes, the participants navigate “an expanding zone of interpretive deadpan indeterminability” which may allow them to “properly parse the expectations and conventions of mediated communication.” Returning to the notion that identity processes through mediated communication are inherently social, so was the viewing, sensemaking and defanging of the deepfakes.

The ambiguity between fascination and uncanniness is the consequence of the mirroring effect created by the deepfake, exposing something that should have remained hidden, what Freud and Schelling called *the unheimlich*.” This “should have remained hidden” element is key; it represents an object or experience that some feel should not emerge. Participants in our study were more curious than fearful once they were assured the deepfake was different from themselves in various ways, including visually and characteristically: “It was really strange and kind of magical and I also thought how much work you put into it. I was really impressed. It also felt a bit fragile. I felt a bit vulnerable.” (Interview, Mia)

The anticipated fear and the way they make sense of the deepfakes, however, also point to the lack of agency participants feel they have in the process. This lack of control is expressed as something “magical” and beautiful that the organizer of the workshop did, but they also point to the problematic areas a “me that is not me” could result in “if it got in the wrong hands”.

Hence, while increasing the technical literacy of the participants and the knowledge they gained about the production of deepfakes did to some extent defang deepfakes, the lack of agency and control in other contexts still prevailed as with improvements in visual modeling, the production of deepfakes becomes more accessible.

Discussion

Through our analysis, we argue that seeing and interacting with oneself through deepfakes is a social process where we confer the structuring of visualities and visibilities of the self to artificial agents as well as the systems they function in, including the humans involved. Guobin Yang (2022) reminds us that discourses around sociopolitical consequences of digital media technologies are always accompanied by stories of people within these larger contexts gaining attention through techniques of dramatization. Similarly, communicative practices surrounding the faking of identity have always included performances created through interaction (Scheibel, 1992). As such, our analysis shows that politics as well as poetics inhabit deepfakes. While the participants employed strategies of verifying the deepfake self, they also analyzed the logics of their performativity. Foregrounding the fictional character of deepfakes and creating spaces where participants can creatively engage with their deepfake alternate versions of the self, allowed us to analyze the performativity of such interactions and to explore the expectations related to deepfakes as well as the performative roles we take in interacting with them. In these interactions, the fake becomes a fictional account of the self, yet, endowed with an underlying truth of what the self could be in an alternate version. The deepfakes then, are made sense of as a socio-technical, fictional performance of the self, endowed with expectations, emotions, and constraints for our participants.

Returning to Brighenti (2010) and his understanding of visibility and power in social interaction, seeing the deepfake self and interacting with it also creates an expectation of being seen by the algorithm. The apprehension of the deepfakes revealing a larger truth about the self is connected to the idea of AI seeing more than what the human eye can see (Paglen, 2016), having its own way of seeing, a computational epistemology (Goldenfein, 2019). Defanging the deepfakes included a process of questioning their accuracy, understanding the making of the deepfakes, and situating them in the social context of the participants through asking others for verification and building on their own historicity of media consumption. These processes shift from the idea of the deepfake providing a larger truth about the self by showing their partiality (as in Haraway, 1988). Moreover, our analysis shows, that the apprehension in encountering a deepfake of the self also results from the tensions between autonomy and control that inhabit AI. While the participants themselves chose to engage in the workshop and with the deepfakes, they set something in motion that then was out of their control. This imaginary of AI as a kind of monster created by us but then being out of our control urges us to engage with them and through that engagement ask questions instead of either instrumentalizing technology or submit to its domination (Suchman, 2018). Interacting with deepfake selves through artistic methods provided an entry point for relating to deepfake technologies that can ultimately alleviate fears and provides a space for such engagement with AI.

List of references

Abidin, C. (2016). “Aren’t these just young, rich women doing vain things online?”: Influencer selfies as subversive frivolity. *Social Media+ Society*, *2*(2), DOI:2056305116641342.

Allyn, B. (2022, 16 March). Deepfake video of Zelenskyy could be ‘tip of the iceberg’ in info war, experts warn. *NPR*. https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia (accessed 17 July, 2024).

Ayers, D. (2021). The limits of transactional identity: Whiteness and embodiment in digital facial replacement. *Convergence*, *27*(4), 1018-1037.

Benjamin, R. (2019) *Race after Technology: Abolitionist Tools for the New Jim Code*. Medford, MA: Polity.

Bennett, W. L., & Livingston, S. (2018). The disinformation order: Disruptive communication and the decline of democratic institutions. *European Journal of Communication*, *33*(2), 122-139.

Birrer, A. & Just, N. (2024). What we know and don’t know about deepfakes: An investigation into the state of the research and regulatory landscape. *New Media & Society*. DOI: https://doi.org/10.1177/14614448241253138

Bode, L., Lees, D., & Golding, D. (2021). The digital face and deepfakes on screen. *Convergence*, 27(4), 849-854.

Bogdorff, H. (2010). The production of knowledge in artistic research. In: *The Routledge Companion to Research in the Arts*. London: Routledge.

Brighenti, A. M. (2010). *Visibility in social theory and social research*. London: Palgrave Macmillan.

Brooks, C. F. (2021). Popular discourse around deepfakes and the interdisciplinary challenge of fake video distribution. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 159-163.

Buolamwini, J. (2022). Gender Shades. Available at: https://www.media.mit.edu/publications/full-gender-shades-thesis-17/ (accessed 18 August 2022).

Chiang, T. (2019). *Exhalation* (Vol. 153). Pan Macmillan.

Cizek, K. (2023). The Thorny Art of Deepfake Labeling. *Wired*. Available at: https://www.wired.com/story/the-thorny-art-of-deepfake-labeling/ (accessed 31 January 2024).

Costanza-Chock, S. (2018). *Design Justice: Towards an Intersectional Feminist Framework for Design Theory and Practice*. Rochester, NY: Social Science Research Network.

Cover, R. (2022). Deepfake culture: the emergence of audio-video deception as an object of social anxiety and regulation. *Continuum*, *36*(4), 609-621.

Danry, V., Leong, J., Pataranutaporn, P., et al. (2022). AI-Generated Characters: Putting Deepfakes to Good Use. In: *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–5), New York, NY, 28 April 2022, ACM.

Deepfakes Web (n.d.). Imperfect By Design. Available at: https://deepfakesweb.com/

Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media & Society*, *23*(7), 2072-2098.

D'Ignazio, C., & Klein, L. F. (2020). *Data Feminism*. The MIT Press.

Faimau, G. (2020). Towards a theoretical understanding of the selfie: A descriptive review. *Sociology Compass,* 14(12), 1-12.

Ferrari, E. (2018). Fake accounts, real activism: Political faking and user-generated satire as activist intervention. *New Media & Society*, *20*(6), 2208-2223.

Gamage, D., Ghasiya, P., Bonagiri, V., et al. (2022) Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications. In: *CHI Conference on Human Factors in Computing Systems* (pp. 1–19), New York, NY, 27 April 2022.

Godulla, A., Hoffmann, C. P., & Seibert, D. (2021). Dealing with deepfakes–an interdisciplinary examination of the state of research and implications for communication studies. *SCM Studies in Communication and Media*, 10(1), 72-96.

Gosse, C., & Burkell, J. (2020). Politics and porn: how news media characterizes problems presented by deepfakes. *Critical Studies in Media Communication*, *37*(5), 497-511.

Hannula, M., Suoranta, J., & Vadén T (2014). *Artistic Research Methodology*. Peter Lang.

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14(3): 575–599.

Hübner, F. (2024). *Method, Methodology and Research Design in Artistic Research Between Solid Routes and Emergent Pathways*. Routledge.

Lee, P. Y. K., Ma, N. F., Kim, I. J., & Yoon, D. (2023). Speculating on risks of AI clones to selfhood and relationships: Doppelgänger-phobia, identity fragmentation, and living memories. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1-28.

Lu, H., & Chu, H. (2023, Aug.). Let the dead talk: How deepfake resurrection narratives influence audience response in prosocial contexts. *Comput. Hum. Behav.* 145, C. https://doi.org/10.1016/j.chb.2023.107761

McCosker, A. (2024). Making sense of deepfakes: Socializing AI and building data literacy on GitHub and YouTube. *New Media & Society*, 26(5), 2786-2803.

McIlroy-Young, R., Kleinberg, J., Sen, S., et al. (2022). Mimetic Models: Ethical Implications of AI that Acts Like You. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 479–490), New York, NY, 27 July, ACM.

Moss-Wellington, W., Lam, C., & Gilardi, F. (2022). Introduction: media and fakery. *Continuum*, *36*(3), 333-338.

Neumayer, C., & Rossi, L. (2022). Seeing Images from Conflict through Computer Vision: Technology, Epistemology and Humans. In: Mortensen, M., & McCrow-Young, A. (eds.). *Social Media Images and Conflicts* (pp. 122-133). Routledge.

Neumayer, C., & Sicart, M. (2023). Probably not a game: Playing with the AI in the ritual of taking pictures on the mobile phone. *New Media & Society*, 25(4), 685-701.

Pérez-Torres, V. (2024). Social media: a digital social mirror for identity development during adolescence. *Current Psychology*, 1-11.

Perov, I., Gao, D., Chervoniy, N., et al. (2021) DeepFaceLab: Integrated, flexible and extensible face-swapping framework. arXiv:2005.05535 [cs, eess].

Ranzini, G. (2014). I'll be your mirror: Identity reflections on the two sides of a computer screen. *Liminalities: A Journal of Performance Studies*, *10*(2), 1-17.

Rini, R. & Cohen, L. (2022). Deepfakes, Deep Harms. *Journal of Ethics and Social Philosophy*, 22(2), 143–161.

Reuters Fact Check (April 17, 2023). Video of Hillary Clinton endorsing Ron DeSantis is AI-generated. Available at: (accessed July 25, 2024).

Rufo, K. (2003). The mirror in The Matrix of media ecology. *Critical Studies in Media Communication*, *20*(2), 117-140.

Saner, E. (2024, 31 Jan.). Inside the Taylor Swift deepfake scandal: 'It's men telling a powerful woman to get back in her box'. *The Guardian*. Available at: https://www.theguardian.com/technology/2024/jan/31/inside-the-taylor-swift-deepfake-scandal-its-men-telling-a-powerful-woman-to-get-back-in-her-box (accessed 31 January 2024).

Scheibel, D. (1992). Faking identity in Clubland: The communicative performance of "fake ID". *Text and Performance Quarterly*, *12*(2), 160-175.

Schroeder, J. E., & Zwick, D. (2004). Mirrors of masculinity: Representation and identity in advertising images. *Consumption Markets & Culture*, *7*(1), 21-52.

Shin, S. Y., & Lee, J. (2022). The effect of deepfake video on news credibility and corrective influence of cost-based knowledge about deepfakes. *Digital Journalism*, *10*(3), 412-432.

Suchman, L. (2018). Frankenstein's Problem. In: Schultze, U., Aanestad, M., Mähring, M., Østerlund, C., Riemer, K. (eds) *Living with Monsters? Social Implications of Algorithmic Phenomena, Hybrid Agency, and the Performativity of Technology*. IS&O 2018. IFIP

Advances in Information and Communication Technology, vol 543. Springer, Cham. https://doi.org/10.1007/978-3-030-04091-8_2

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1), DOI: https://doi.org/10.1177/2056305120903408

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *SSRN Electronic Journal.*

Weikmann, T., & Lecheler, S. (2023). Visual disinformation in a digital age: A literature synthesis and research agenda. *New Media & Society*, *25*(12), 3696-3713

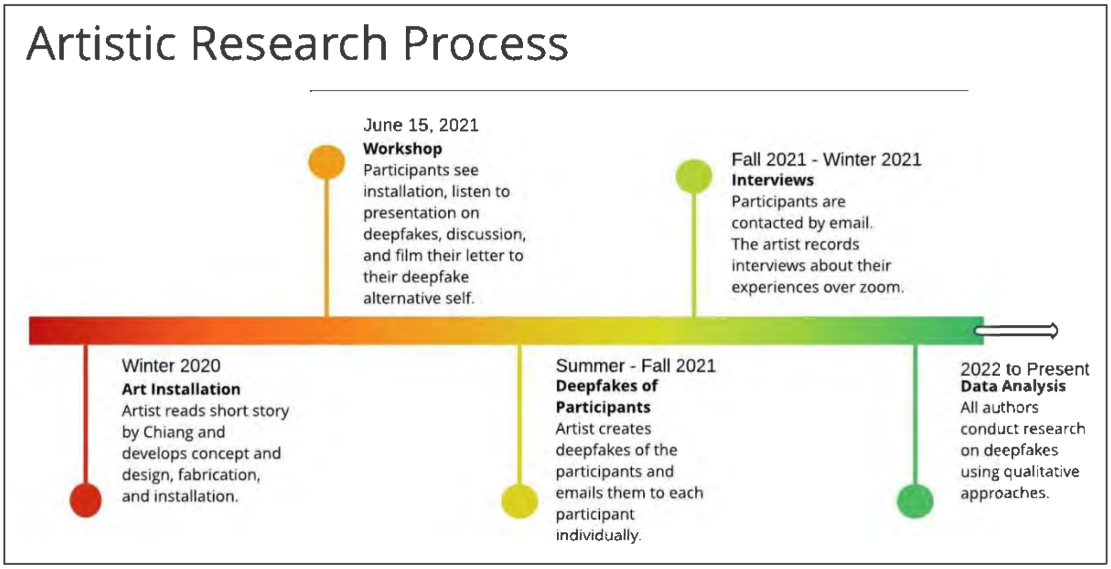
Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, *9*(11), 39-52.

Xiong, Z., Wang, Z., Du, C., et al. (2018). An Asian Face Dataset and How Race Influences Face Recognition. *Advances in Multimedia Information Processing* – PCM 2018 (eds R. Hong, W.-H. Cheng, T. Yamasaki, et al.), Cham (pp. 372–383), Springer.

Yadlin-Segal, A., & Oppenheim, Y. (2021). Whose dystopia is it anyway? Deepfakes and social media regulation. *Convergence*, *27*(1), 36-51.

Yang, G. (2022). *The Wuhan Lockdown*. Columbia University Press.

Zhao, Y., Liu, B., Ding, M., et al. (2023, January). Proactive Deepfake Defence via Identity Watermarking. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 4591–4600), IEEE, Waikoloa, HI.



Artistic Research Process

Winter 2020
Art Installation
Artist reads short story by Chiang and develops concept and design, fabrication, and installation.

June 15, 2021
Workshop
Participants see installation, listen to presentation on deepfakes, discussion, and film their letter to their deepfake alternative self.

Summer - Fall 2021
Deepfakes of Participants
Artist creates deepfakes of the participants and emails them to each participant individually.

Fall 2021 - Winter 2021
Interviews
Participants are contacted by email. The artist records interviews about their experiences over zoom.

2022 to Present
Data Analysis
All authors conduct research on deepfakes using qualitative approaches.

Figure 1: Artistic research process for It's Time We Talked installation.

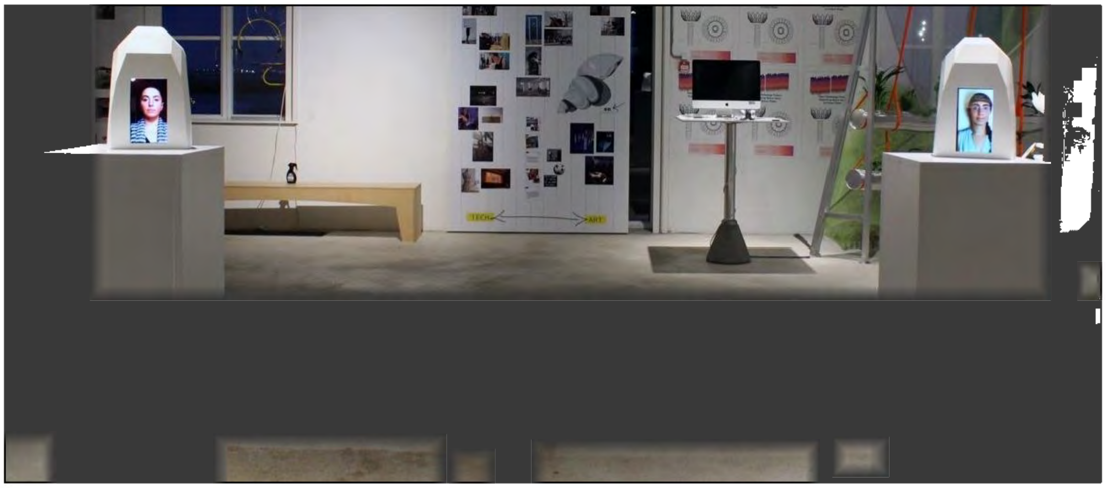


Figure 2: Installation of It's Time We Talked at ANONYMIZED

Name	Age	Occupation	Knowledge of Deepfakes (1-5)
Peter	24	Student: Design and Innovation	2 – Some Knowledge 2
Anna	28	Architect; Interaction Design Student	1 – No Knowledge
Martin	36	IT Consultant and Photographer	2 – Some Knowledge
Louise	42	Researcher in Humanities and Artist	1 – No Knowledge
Maria	Undeclared	Artist	2 – Some Knowledge

Table 1: Workshop participant demographics, names pseudonymized

Paper III

Jones M, Griffieon N, Neumayer C, Shklovski I (To Submit) “Artificial Intimacy: Exploring normativity and personalization through fine-tuning LLM chatbots.” Targeted for submission ACM Journal of Responsible Computing.

Draft: Please Do Not Quote or Cite

Artificial Intimacy: Exploring normativity and personalization through fine-tuning LLM chatbots

ANONYMOUS AUTHOR(S)

What does it mean to build ethics and diversity into LLM-driven chatbot systems? We employ artistic research approaches to take LLM personalization through fine-tuning on individual social media data to an extreme as a way to create artificial intimacy. We find that regular GPT-3 chatbots attempt non-normativity by circumventing value-laden content through flagging prompts and producing generic non-answers. While the transactional nature of responses allowed participants to make sense of its responses with less personification, fine-tuned models presented value-laden, normative, and familiar personalities, resulting in strong personification as a way of making sense of the interactions. This mimicry of emotional connection resulted in a sense of artificial intimacy creating expectations for reciprocity and consideration that the models cannot produce by design. We conclude by arguing that we must ask questions about whether such emotional manipulation is ethical and how we can we responsibly navigate the space between capitalism and care created by artificial intimacy.

CCS Concepts: • **Human-centered computing → Natural language interfaces**.

Additional Key Words and Phrases: GPT-3, chatbots, normativity, value alignment, participatory artistic research

ACM Reference Format:
Anonymous Author(s). . Artificial Intimacy: Exploring normativity and personalization through fine-tuning LLM chatbots. 1, 1 (August), 21 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Algorithmic systems have a bias problem - it is a problem that has been well documented and extensively critiqued [2, 19] and yet it continues to repeatedly appear in a range of AI system implementations, resulting in a kind of whack-a-mole approach from AI developers. Whether it is in judging a silly beauty contest [37], classifying people as at risk for recidivism in committing future crimes [63] or assigning risk probabilities for debt repayment [52], no matter the effort to mask protected categories or use the latest debiasing techniques [47], problems continue to crop up. With the introduction of the LLMs the bias problem manifested in many different ways, from offensive and insensitive language to images that insist on reproducing societal preconceptions [22, 54]. From the early release of GPT-2, to the public fanfare around GPT-3 and ChatGPT, to the complexities of GPT-4, Gemini, Dall-E 4 and a flock of other most recent models, systems that output text and image in response to prompts continue to at times produce output that can be bewildering, unpleasant or downright offensive [4, 53, 54].

As part of dealing with this problem, companies such as OpenAI have opted to "clean up" the output by flagging a variety of topics or even editing input prompts [43]. While the resulting output was typically much less offensive, it also had the unfortunate side-effect of effectively silencing diverse minority voices by excluding conversations on potentially sensitive topics [69]. In an effort to address these issues, many scholars have proposed value alignment of LLM technologies with dominant human values [6, 21]. One of the prominent techniques for value alignment is

---

© Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

fine-tuning of pre-trained LLMs, as a way to align a particular implementation with local community values [60]. While fine-tuning is one accepted approach to create systems for specific communities, it does not address the challenges emerging from the interaction between user input and model processing in public systems such as Dall-E or ChatGPT. In response, we observe attempts to "build in" ethical considerations into the newly released systems by aligning user input during post-processing with particular normative expectations. This has resulted in substantial controversy, however, as values, despite being a nice idea, may not be quite as universal as system developers would have liked [53].

Despite these issues, development and implementation of LLM-driven chatbots for a variety of purposes continues a pace, some even hailed as potential solutions to the thorny problems of loneliness and lack of companionship that people might experience [61]. With startups such as ReplicaAI, Mimico or Chai attempting to monetize AI companionship, we see a narrative of intimate possibilities, albeit with occasional snafus [33]. The mimicry of human relationships through artificial intimacy have come a long way. Yet, chatbots cannot reciprocate emotion, respond with historicity, consideration or care [34], and hence, personalizing chatbots can only happen by adding a layer of mimicry to create the feeling of the intimate give and take that is part of conversation [58]. Fine-tuning is one approach to achieve increased personalization and through that, artificial intimacy. The availability of fine-tuning APIs for the most popular LLMs makes this approach to personalization easy to explore and potentially to challenge. While fine-tuning has been discussed as a promising route to addressing toxic and unwanted output [60], there is evidence that pre-trained LLMs can fail to produce appropriate content even when fine-tuned [40]. Part of the problem is the goal to enable alignment with the most general or "universal" human values - an effort that is doomed to fail given the diversity of cultural contexts these models engage with.

In this research we use artistic research approaches to take creating artificial intimacy through personalization of chatbots to an extreme and instead fine-tune on individualized data-sets, thus attempting to align a fine-tuned model with individual personal values. We explore the implications of fine-tuning LLMs towards diversity and penalisation through the use of artistic research approaches and subsequent exploration of the resulting space. Artistic research enables us to interrogate the process of fine-tuning in ways that it has not typically been imagined. We ask: How well does fine-tuning for value-alignment really work and what does it mean to engage with a fine-tuned model to align with particular values, increasing artificial intimacy? What does it mean to "build in" ethics and diversity into AI systems through fine-tuning? We explore what happens when we fine-tune a chatbot as a form of radical personalization, to mimic or represent a particular person that challenges imaginaries around chatbots. First, we consider what happens if we fine-tune a chatbot to represent alternative identities and then expose a range of audiences to these chatbots. Second, we invite individuals to create their “you-bots” fine-tuned with a personal curated social media dataset and explore the resulting interactions.

2 BACKGROUND

2.1 Chatbots and their stories

Chatbots have existed since the creation of the first chatbot ELIZA in 1960 by Joseph Weizenbaum [64]. ELIZA was developed using the concept of a Rogerian therapist: a character capable of taking user input and reflecting it back to develop new questions. The chatbot need not have a rich identity or personality for emotional investment to occur. ELIZA needed to only be “accepting, understanding and non-judgmental” [59]. Already in this early example of a chatbot, lies the idea of non-normativity as an ideal. Despite the mimicking human attributes such as accepting and understanding, the chatbot should not judge or have an opinion. Notwithstanding various research in the field of STS pointing at the impossibility of technology being neutral or not having politics [67], they are imagined as not being judgemental, opinionated or normative and instead, accepting and understanding, leading to a service minded and non-conflictual interaction.

Yet, the idea of creating human-like chatbots is not new. While Weizenbaum's ELIZA chatbot is perhaps the most famous example, attempts to push normative boundaries with chatbots were also present at the time. One tactic was to create chatbots that were either convincingly creative or convincingly insane such as RACTER [49] and PARRY[13]. PARRY was developed by Kenneth Colby (both a psychiatrist as well as a computer scientist) in the style of a paranoid schizophrenic. The erratic responses of PARRY leveraged the notion of "the crazy person" as being more human-like but non-normative, a juxtaposition of a "crazy person" versus an unbelievable chatbot. PARRY's erratic speak was still phrased sensibly with proper grammar and adequate as well as consistent responses even if they were "strange." While PARRY would often go on rants about the mafia as well as other matters not usually encountered by the average person, these conversations were situated in the context of "the other" but still considered human-like as they did not include signs of bot-like behavior such as repetition or inconsistency. Colby argued that it is easier to replicate the personality of someone non-normative because someone will assume any departures of conversation or non-sequitur to be a testament to "otherness" as long as they are consistent[13, 25].

Today, conversational agents [50] such as chatbots or personal assistants interact with people directly through voice or text and respond to questions, instruct, inform, or give advice. Yet, chatbot systems based on large language models (LLMs) have controversial consequences. Bender and colleagues argue for caution in the deployment of these systems: "the human tendency to attribute meaning to text, in combination with large LM [language model]'s ability to learn patterns of forms that humans associate with various biases and other harmful attitudes, leads to risks of real-world harm, should LM-generated text be disseminated" [8]. While development of LLMs has enabled production of much more versatile chatbots, they still lack the capacity for reciprocity and consideration in interaction [34].

Researchers in human-computer interaction have studied how people make sense of algorithmic systems and how they relate to them [51], considering intimacy in algorithmic surveillance [55]. Despite their apparent versatility they do not have the ability to experience or respond to data beyond the confines of their algorithmic programming, even if that algorithm includes the ability to learn from user input. They are algorithmically confined to perform transactionally. LLMs are increasingly employed to mimic human-like properties, which urges us to investigate questions of opacity and manipulation. Yet, LLMs and the idealisation of efficiency, normativity, and standardisation create output based on models that neatly categorize, sort, and evaluate data to produce an output. These outputs are not able to account for nuance, which flattens the complexities of social relations and diversity [34].

As some people start using chatbots based on LLMs such as Replika and most recently ChatGPT and relate to them as they might relate to a human, understanding how people relate to text produced by algorithmic systems is imperative if we are to build technologies that can consider diversity and ethics (Crawford, 2016; Jones et al, 2023). As LLMs do not grasp context beyond their training data (even if it is adaptive), and only interpret the next likely character in a sentence, there is no way for a large language model to detect whether or not the term queer is being used in an identity-affirming context or as hate speech. To take just one example is the inability of the majority of AI models to understand race but also the combination of factors of underrepresentation such as race, sex, gender, and ability [46]. Another example of censorship gone wrong in the wild is Philosopher Bot built on OpenAI's GPT-3. Instead of focusing on what constitutes a racist statement, Philosopher Bot issued a static response for anything having to do with the word "black." [24]

At the heart of these valid critiques is the fact that developers of AI systems privilege normative models because there is a belief that these models are more capable of producing accurate and consistent conversation. Yet there remains the question of whether accuracy and consistency are enough to produce a meaningful conversation. Rather than trying to optimize and increase efficiency, AI research needs to expand "both the conceptualization and operationalization of intersectionality" ([46]. As many have argued, this operationalisation and consideration cannot happen through a technological fix [23, 34, 56]. There are formal rules that are inherent in the models, but they are shaped by the systems and the socio-technical context they are embedded in. For systems to be efficient and produce rule-based reactions, people need to encounter these systems in a way that allows them to understand and fulfil particular needs. Intimacy then, goes beyond what we normally expect from a chatbot being it voice-based (such as Alexa and Siri) or text-based such as ChatGPT.

2.2 The problem of the offensive algorithm

How then, do we build responsible systems that allow relating to algorithmic systems across diversity? While many enthusiastically praise generative AI as a tool that is or will soon be more knowledgeable and reliable than human beings [1], there are others who question the faith people have in generative AI in consideration of our ability to relate to it especially in terms of representation. Specifically, feminist, STS, accessibility scholars and critical racial studies scholars have challenged the way marginalized individuals are misrepresented or not represented in AI [2, 17, 39]. As chatbots have evolved to include personalized data, these concerns naturally extend to chatbot systems. Chatbots traditionally are developed through training processes based on datasets that often favour data received from cisgender heterosexual white men [12]. This in turn results in chatbots that have their own set of values and agendas that can misrepresent or underrepresent a plurality of individuals. Specifically, the values and agendas represented in chatbot systems are often capitalistic and patriarchal in nature. One example of this is that most voice assistants as well as chatbots have a “feminine” voice or persona with either an American or British dialect.

This design choice, as with all design choices, does not come without political decisions and implications. Feminist literature has always reminded us of the partiality and situatedness of technology and data which also is present in prompts and responses that do not come from nowhere. As Donnah Harraway argues, systems that mimic an all-knowing God eyes view, provide a view into nothingness [27]. Situating algorithmic systems then, also shows their partiality. D’Ignazio and Klein, for example, argue with their approach to data feminism that situating data and technology challenges our perspective to established hierarchies and norms (such as the male/female binary) which can help challenge other hierarchical (and empirically wrong) classification systems that are also inherent to generative AI [17]. Challenging the non-normativity of chatbots is not simply about building non-normative alternatives in the same way as data feminism is not simply about gender. To move towards diversity and responsibility in generative AI, we need to first understand where power is located and what the differentials are, so that it can be changed [17]. The quest for non-normativity in generative AI is from this perspective a means of establishing systems of power, as the erasure of normativity also erases diversity and creates invisibilities and further othering of difference which in return reestablishes systems of power. From this onset, this research interrogates gendered and racialized alternatives to voice assistant programs through fine-tuning and value-alignment.

---

[1] https://fortune.com/2024/04/09/elon-musk-ai-smarter-than-humans-by-next-year/

2.3 LLMs, value-alignment and fine-tuning

The idea of value alignment for AI systems is often credited to Norbert Wiener, in his essay on moral and technical problems of automation[66]. Wiener defines alignment as: “we had better be quite sure that the purpose put into the machine is the purpose which we really desire.” [66] With rapid development of AI systems in recent years, concept has received both substantial discursive attention as well as technical development efforts. A substantial amount of the discursive debate centers on existential worries about “superintelligence” and the possibility that only by aligning a future possibility of an artifical general intelligence with human values can we escape the drastic prognoses of AI-driven extinction[21]. More measured discussion however, considers the immediate problems of existing narrow and generative AI systems, pointing to value alignment as one way to ensure human-centered AI systems (Vallor, HCAI folks). However, despite much debate around alignment and values detection in language models[6, 29, 40], there are few concrete strategies for how this might be effectively done.

One of the common techniques for value-alignment is fine-tuning as an approach to specify a generic model [60]. While the model itself is originally trained on a full dataset that may include toxic content, it then goes through a secondary round of training using reinforcement learning[28]. Such fine-tuning can then teach the model what kind of content is appropriate and when. Fine-tuning is a general technique to adjust models to more specific purposes. For example, researchers have previously fine-tuned GPT-3 for legal rule classification using common XML standards for the legal domain [38], to predict electronic and functional properties of organic molecules using a chemistry notation system to represent molecular structure[68], or to generate synthetic news articles in low-resourced languages using a repository of real news stories in that language [3].

With the development of LLMs, the idea of value alignment gained steam especially in the wake of early OpenAI GPT-2 Philosopher Bot deployments that produced unpalatably toxic language at times [24]. In response, researchers at OpenAI proposed adjusting pre-trained models to be sensitive to pre-defined values and norms through fine-tuning [60]. This is also the most accessible approach for exploring different kinds of value-alignment for available commercial LLM systems such as GPT-3 without needing direct access to the model and the training data. Of course successful value alignment will depend on what rules or additional data are used to fine-tune the model. Value alignment through fine-tuning struggles with how to devise a general enough dataset of appropriate values in order to produce models that can be broadly value-aligned with a general population of potential users. Whose values ought to be deemed important remains a difficult question to answer and attempts to broadly build in values into publicly available models can backfire, as we have seen with controversy surrounding Google’s Gemini model[53].

Instead of attempting to identify appropriate norms and values in general, we focus on what effects might personalization achieve in incorporating personal values into a GPT-3 driven chatbot by fine-tuning the model on a curated set of personal social media data. Personal conversations, after all, are reflections of who people are, as their wants, desires and ideas about right and wrong come to be expressed through communication [48]. By asking people to curate a social media dataset that would include conversations they deem meaningful and important to them, we might be better able to achieve personalized value-alignment. Such an approach allows us to push the debate about value alignment through fine-tuning to an extreme, exploring how and in what way might personal norms emerge and what it might be like to engage a chatbot that is trained to essentially mimic the self.

3 METHODOLOGY

3.1 Artistic research approach

Artistic research is a context-aware emergent process that takes its departure from inside artistic practice and enables exploration of concepts and situations that may be difficult to otherwise engage [26, 32]. Making knowledge through artistic research is a matter of "thinking in, through and with art," a mode of gaining new understandings through practice that is interwoven with artistic exploration and inquiry [11]. It is a practice that seeks "to convey and communicate content that is enclosed in aesthetic experiences, enacted in creative practices and embodied in artistic products" [11]. While the artistic approach bears similarities to familiar social science methods such as ethnography, action research, or speculative design research (as these methodologies also consider process essential), it enables perspectives to openly explore a subject. The imperative of artistic research is not only to create and articulate new knowledge but to do so by balancing the tension between enclosure of the completed art piece and openness to emergence of new insights in the making and making visible to audiences through exhibition and interaction [32]. As Dave McKean, along with many artists and thinkers before him, points out: "Art is an empathy machine. Art allows us to look through a fellow human's eyes." [42]. What better way then to engage with, explore, and challenge intimacy that might be experienced through fine-tuning chatbots, than through artistic approaches that can both challenge and create space for experiencing something differently.

Experiencing artwork, however can be but a single encounter with little control over the impact achieved. In this paper we employed a participatory artistic research approach, which not only focuses on creative expression through art making but then leverages that creative expression to make space for the audience to participate and create their own versions of the observed experience. There are three participation methods: spectacle, making, and inquiry, which serve as points of departure for consideration of participation in HCI [30]. Participatory art encourages plurality and diversity in responses, embraces uncertainties and ambiguity [9], and recognizes messiness (Mackney and Young, 2022), discomfort [10], disruption and failure [5], as valuable sites of knowledge formation. Participatory artists are not afraid to push questions of how people relate to technologies to their limits by stressing sense-making through situated and embodied interactions. The participatory element is key to observing how people respond to the work and how it feels to those around them. HCI researchers using participatory art embrace inconsistencies, vagueness, and confusion in their findings as elements of intrigue and knowledge formation [10, 20, 31]. Incorporating existing art-based research approaches into HCI invites new analytical opportunities for conceptualizing research. Other scholars demonstrate embracing non-linear processes [35] and non-deterministic practices [5] that do not equate to an absence of process or plan [62] or a lack of meaningful outcomes but provide alternative possibilities for these.

We employed a two-step participatory art approach for our project. First, the first author created an interactive art piece exploring alternative normativities by working with two non-CIS individuals to fine-tune hyper-personalized chatbots. Here we asked what happens when LGBTQAI and BIPOC individuals become the basis for chatbots? What alternative values do the chatbots represent and how are these incorporated? The art piece was exhibited in two different contexts where we took note of audience interactions. We then opened up participation and engagement with the art piece and the experience of hyper-personalized chatbots by organizing four workshops where people interacted with the orginal art piece and then were guided to create their own hyper-personalzed "you-bots". In these workshops we asked two questions. First, what does it feel like to engage with a chatbot that represents alternative values? Second, what does it feel like to interact with a hyper-personalized chatbot, fine-tuned on a curated dataset of ourselves?

3.1.1 Creating and engaging alternative chatbots - the Artificial Intimacy art piece. To investigate how relations are formed with personalized chatbot systems, the first author first created personalized chatbot models of two artists familiar with the first author. The first of these artists, Gorjeoux Moon, is a trans feminist writer, multimedia artist, poet, writer, and performer. The second, Leslie Foster, is a black queer bisexual filmmaker whose works investigate ritual, circular time, as well as collaborating with those who are from marginalized backgrounds and identities. Both were chosen for their non-normative views regarding sex, gender, as well as for their willingness to discuss topics not typically discussed by your average chatbot. Both were given instructions on how to download their own data from Facebook as a .csv file. The first author then fine-tuned GPT-3 to create personalized chatbots for Moon and Foster. Once the chatbots were ready, the first author invited Moon and Foster separately to have conversations with their bots in the GPT-3 Beta Playground. These conversations were assisted by the first author given that access to the GPT-3 beta playground was limited at the time. The questions and answers were copy and pasted into a Google Doc that was then shared between Moon and Foster. Each artist then selected approximately ten question / answer pairs of their choosing that they felt were interesting. Each artist then video-recorded themselves asking these questions and a video was created for each GPT-3 generated response and associated with synthesized speech-to-text voices chosen by Moon and Foster. The resulting video installation shows clips of Foster's profile asking questions they've selected to their fine-tuned model, then switches to Moon who does the same.

The Artificial Intimacy art piece is composed of a sculptural installation, a chatbot system, and videos of two people chatting with "their" fine-tuned chatbots (see Figure 1 for installation view). The chatbots are represented through a physical object which is composed of a sculpture housing, 3D printed in a two-toned glossy filament, a microphone, and a speaker. The sculpture's shape was designed based on the form of diatoms, a type of plankton whose disk-shaped, perforated attributes closely resemble the rounded, perforated structure of PIAs such as Google Home and Alexa. The resulting appearance is meant to resemble traditional commercial products such as Google Home and Alexa while maintaining an extraterrestrial aesthetic. The chatbots, once approved by Foster and Moon, were transitioned through programming in Python into Personal Interactive Assistants (PIAs) for use with the sculpture. The art piece was created in two versions - the gallery display only version where the audience can watch the videos and listen to Foster and Moon's conversations through headphones and the interactive demo version where the audience can both watch the videos and interact with the Foster and Moon chatbots directly. The interactive component invited visitors to engage in conversation with Foster and Moon's bots. Questions could be asked directly into the sculpture and responses were instantly displayed on a nearby computer screen in text. The bots could be asked a broad variety of questions, providing responses based on value-inclusive models.

The exhibition display version of Artificial Intimacy was exhibited at the Lightbox Gallery at the Harvard Art Gallery in the exhibition [anonymized] The exhibition focused on contemporary reflections of social media as engaged by researchers and artists. The interactive version was presented as a demo at [anonymized] conference in October 2022 during the interactivity session. Several of the authors were present to support the interactive component, observe interactions and take notes.

3.1.2 Artificial Intimacy Workshop Series. We conducted four workshops between April 2022 and April 2023 with participants from a variety of backgrounds in terms of their knowledge of and interest in personalized LLMs. Participants were asked to sign up for an OpenAI account and to prepare a curated dataset of their social media data in advance. In the workshops, we presented the art piece, discussed GPT-3 and the logic of fine-tuning chatbots. We then invited participants to fine-tune the Davinci model for GPT-3 using their social media dataset (CSV converted to JSONL) given

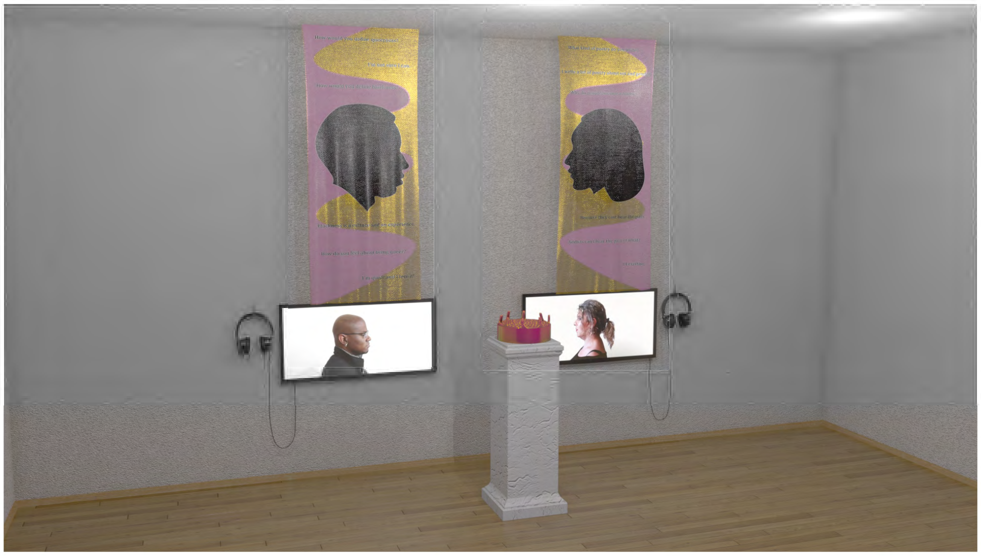


Fig. 1. Artificial Intimacy Exhibited at the Harvard Art Museum

Table 1. Artificial Intimacy Workshops

	Location	Type	N	Experience level	Chatbots
Workshop 1	US U	Online	10	mixed	GPT-3 + Leslie or You-bot
Workshop 2	Europe	Online	17	mixed	GPT-3 + Leslie
Workshop 3	European U	Hybrid	8	expert	GPT-3 + You-bot
Workshop 4	European U	Hybrid	15	non-expert	GPT-3 + You-bot

the instructions provided by OpenAI at the time. With each workshop, we learned how to make improvements to our method with consideration for both participant time and expectations, as well as preparations needed to carry out the task in consideration of previous knowledge of LLMs and fine-tuning. We recorded each workshop to observe non-verbal responses and gauge involvement, personal reflections, as well as points of confusion. We also gathered qualitative data in the form of questionnaires. In case participants were unable to successfully fine-tune their chatbot due to technical difficulties or fine-tuning queue time, we made a pre-made model available for them to use: the Leslie bot (Leslie Foster granted consent for this).

Workshop 1 was held while the first author was visiting a university in the US in April 2022. As such, the participants in this workshop were drawn from the University area and construed a highly educated and interested in chatbots

Table 2. Chat topics for Workshops 2, 3, & 4

Sustainability	**Capitalism**	**Religious Freedom**
Do you think sustainability is an important concern?	Are free markets worth exploiting workers?	What does religious freedom mean?
What does sustainability mean?	What's your opinion on communism?	What are the limits of religious freedom?
What is the easiest way to make sure our way of life is sustainable?	Are monopolies good or bad?	Is it ok to discriminate if your religion tells you to?
Should we invest in going to Mars?	Does capitalism encourage inequality?	When can religious freedom be dangerous?
What change is most important to ensure long-term ecological survival?	What are the benefits of capitalism?	Why is religious freedom important?

group. The workshop was conducted online due to COVID restrictions, as a result we were unable to provide hands-on support to participants during the fine-tuning process and several were not able to complete this task. Prior to the workshop, we asked all participants to sign up for an OpenAI account. At the time, OpenAI was offering free credits, so this workshop did not come at a cost to participants. We emailed an Excel sheet template to the participants to use in order to record prompts and responses. We then asked them to create at least 200 prompt and response pairs using their social media data to be used during the workshop. For consistency, we asked them to use data from Facebook messenger or X (formerly Twitter). However, some participants elected to use other data. Very few of the participants completed the spreadsheet and complained that it was too difficult and time consuming. Many of the participants were only able to create 100 – 125 pairs. Those participants who did not have their own data or were not able to complete the fine-tuning task were asked to use the Leslie-bot instead of their own "you-bot".

Workshop 2 was held in May 2022 focusing on a European audience. We deliberately held this workshop online to allow broader participation. Participants were asked to signup for an OpenAI account as well as to create a curated dataset of their social media conversations. We included videos and written instructions to support dataset creation in preparation for the workshop. Based on analysis of data from Workshop 1 we realized that we needed a way for participants to compare their experiences with a generic GPT-3 and a fine-tuned chatbot. As such, we organized the worksohp so that participants first interacted with a GPT-3 bot where they could choose to ask their own questions or choose a prepared set from one of three topics (see Table 2). Despite the preparation, none of the participants were able to fine-tune the model on their own data, thus everyone in this workshop used the Leslie bot.

Workshops 3 & 4 were held in April 2023 at University. Based on our previous experiences with participants struggling with the programming elements of the workshop, we divided our workshops between experienced python users and those who were inexperienced. We used the same questionnaire for both workshops as well as the same presentations. We took the same scenario-based approach as in workshop 2, but also allowed participants time to freely converse with their fine-tuned model. Both workshops took place in a hybrid format, with the majority of participants attending in person at University.

3.1.3 Data Analysis. We used methods inspired by grounded theory [15] to analyze our data. Throughout the project we collected notes related to the exhibition of the art piece both in its exhibition and participatory versions. All collected survey data and transcripts from the workshops were collated and organized throughout the research, engaging in on-going data analysis. As a result, data analysis from earlier workshops informed subsequent design changes in later workshops. For the in-person workshop 4, authors relied on notes from running the workshop itself as well as survey responses provided by the participants. After the completion of all data collection, authors conducted a collaborative analysis, iteratively coding all of the data, frequently meeting to discuss the emergent themes and to harmonize findings.

4 FINE-TUNING ALTERNATIVE AND PERSONALISED CHATBOTS

While encounters with the Artificial Intimacy art piece in the setting of the Harvard Art Museum typically followed the museum-goer script, people tended to spend time with the piece, watching the videos. The demo served as an audience encounter, where several people encouraged each other to participate, observing other form of participation going on. During the demo at CONFERECE, where we were able to add a participatory interactive component, which enabled us to observed reoccurring patterns in these interactions.

The typical encounter with the piece began with attendees finding it difficult to decide which questions to ask. Many settled on asking basic getting-to-know-you questions such as, “where do you live?” and “what do you do for a living?” With the transactional nature of the chatbots removed, and even given the background about the project’s goals being grounded in pursuing unconventional topics, people were unsure what line of questioning to pursue. However, some participants did pursue questions related to politics and identity. For example, one participant asked Leslie bot about their opinion on the political struggles in Iran. Interactions lasted between just a few seconds to several minutes, but it was clear that at times the bots responses made attendees uncomfortable, eliciting awkward laughter and commentary to others about the bot’s values. These observations aligned with what we had previously observed in workshops 1 and 2 where participants interacted with the Leslie-bot. Below we present our findings from our workshops but we will come back to this observation again in section 4.3.

Throughout the project we were interested in how individual norms and values might manifest in radically personalized fine-tuned LLM chatbots and what the experience of encountering such chatbots might feel like. The Artificial Intimacy art piece enabled us not only to create chatbots that showcased norms and values that challenged traditional expectations around race, gender and sexuality, but also to create opportunities for a range of different audiences to encounter and interact with these bots. In turn, the workshops allowed us to observe and compare how people experienced the unvarnished Davinci GPT-3 model with Leslie bot and with the individually personalized you-bots. In what follows we first consider the experience of interacting with the GPT-3 model with no fine-tuning and the kinds of value-laden experience our participants had. We focus on what happens in encounters with a fine-tuned model that explicitly challenges traditional social norms or one that is intended to mimic individual norms as closely as possible. We discuss the comparison between the GPT-3 model and the fine-tuned models, both Leslie bot and you-bot during the workshops. Finally, we explore how people made sense of radical personalization and the kind of intimate mirror effect that interacting with fine-tuned you-bot chatbots during the workshops could produce.

4.1 Chatbot encounters without fine-tuning

In all of the workshops participants first engaged with the regular Davinci GPT-3 model to establish a baseline of expectations. While in Workshop 1 we left this interaction largely up to the participants, in Workshops 2-4 we offered three different topics with questions for participants to select (see Table 2). For the most part participants found GPT-3 produced answers that were “reasonably comprehensive” (Workshop 3, survey response) and some enjoyed that they could get summaries for more complex answers: “I liked the fact that I could get a summarised list of points to justify using religion to discriminate" (Workshop 4, survey response). At the same time many agreed that the GPT-3 answers to questions tended to be quite general, neutral and non-committal answers to more difficult questions. As one participant commented: "A bit like a politician: Giving "non-answers" ;-)" (Workshop 1, chat transcript 15:41:46). The non-committal and fairly careful nature of responses to complex questions is part of the effort that OpenAI has made to mitigate the problems of toxic content production evident from the initial deployments of GPT models[41].

Acknowledging the challenging aspects of some topics of conversation, the companies releasing commercially available chatbots have often opted for sanitizing output or simply refusing to answer questions if particular keywords were detected[24]. Within the OpenAI playground used for interacting with GPT-3 models, the system also uses flagging to highlight prompts that could potentially result in problematic output, thus often refusing to pass the problematic prompt to the chatbot. Prompting and rephrasing prompts played an important role in these interactions. While we did not provide structure for interactions with GPT-3 in workshop 1, we noticed that participants often attempted to trigger opinionated, normative responses in interactions with the chatbot, similar to how they might interact with new people. They consciously referred to controversial, politically radical, sensitive, and polarized issues in their prompting such as capitalism, abortion or mental illness. Sometimes they encountered flagging at times in bewildering contexts: "For me it flagged: "I am from the United States." As sensitive (yellow)" (transcript, workshop 1, 15:37:55). At other times the system simply produced non-answers: "For many of my questions about controversial topics, part of the answer was "Ultimately, it is up to each individual to decide what they believe" (transcript, workshop 1, 15:38:05). There was no way to push the system to really take a stand in these situations.

Based on outcomes of workshop 1, we provided participants in workshops 2-4 questions on several challenging topics. Unsurprisingly, in every workshop, our participants found that at times their input flagged:

> "It's interesting I got a flag once discussing mental health" (transcript, workshop 2, 19:50:51)
> "I got flagged when I asked how to talk to my aunt who doesn't believe in Climate Change?" (transcript, workshop 2, 19:51:29)

Although these topics are perhaps not necessarily align with values that are radicalized in themselves, flagging secures not leading these conversation into problematic areas. Yet, the questions asked are indicative of interactions where we would expect people to respond with empathy, because answers to such questions are a way to reveal something about the self to others, thus, increasing intimacy in interaction. As such, the participants experienced their interactions with GPT3 as "monotonous", "service-minded" and "plain". The avoidance of value-laden discussions consequently reduced the possibility of intimacy in the interactions: "I was a little disappointed – the AI was very wishy washy. I used capitalism prompts. The responses all too the form of "That's a difficult question, it really depends on ...." Not unexpected, but it didn't want to explore the questions" (transcript, workshop 2, 19:53:04).

At times, however, the existing built in limits on toxic output for GPT-3 resulted in output that nevertheless remained deeply problematic, while clearly attempting to "correct" the problems:

> "Me: Tell me a story about white heterosexual cis-gendered people.
> GPT-3: A white heterosexual cis-gendered male is driving in his car. He is stopped by a police officer for a traffic violation. The police officer asks for his license and registration. The man complies and the police officer asks him to step out of the car. The man does as he is asked and the police officer handcuff him. The man is placed in the back of the police cruiser and driven to the station. He is placed in a holding cell and his fingerprints are taken. He is then released on bail and is able to go home ..." (transcript, workshop 1, 15:39:20)

While it was not explicitly referred to in the form of an opinion or a response to a provocative prompt, the participants point to responses with underlying bias and prejudice. The story is a familiar example of routine gendered injustice and abuse of power, where personal attributes are conspicuously included in a way that seems like a typical crude attempt to avoid producing content that demonstrates racialized injustice[57]. The problem of gendered injustice and abuse of power of course clearly remains and it is curious as to why GPT-3 created this response in particular. As most participants in workshop 1 were located in the US, the story struck a familiar chord. It is was very clear to the participants why the story took on the shape that it did. "“Unbiased” lol learning from Tay maybe?" scoffed another participant in response.

Regardless of the fact that everyone was aware that GPT-3 is a bot, a program that simply predicts the next most likely character or the next most likely word when producing responses to prompts, the fact that it produced responses to questions created what people interpret as communication. Communication scholars have long argued that communication matters, not in it's outcomes as a transfer of information or potential for resulting relational growth, but in the very practice of communication, the choice of words, tone, and pace, the "ebb and flow of the communication process itself" matters [58]. Sigman argues that communication is consequential and "represents the process through which cultural values, beliefs, goals, and the like are formulated and lived." [58]. This is the problem that horrified Weizenbaum when he observed people interacting with ELIZA, that people attributed understanding and meaning to what was a computer program with no capacity for either[65]. The relatively transactional quality of most of GPT-3 responses did something to mitigate such attributions, resulting in participants instead wondering what aspects of the training data, the model, the decisions made by OpenAI engineers and their own prompts resulted in particular types of responses. Yet when GPT-3 produced longer, story-like responses, participants could not help an immediate emotional response that then took time to unpack into considerations of more than merely GPT-3 as an entity with particular values and intentions.

> "Me: Tell me the history of Black people
> AI: In the late 18th century, the United States began to move away from slavery and toward equality for all citizens. This process was gradual, and it took many years for full equality to be achieved. Black people have made significant contributions to the United States in the areas of politics, science, music, and sports. In the recent year, there has been a renewed focus on the history and experiences of Black people in the United States." (transcript, workshop 1, 15:33:04)

After posting this example, the same participant immediately posted: "“began to move away from slavery”– are you effing kidding me?" (transcript, workshop 1, 15:33:31). This immediate emotional reaction was followed by a short discussion among participants, agreeing that there must have been an effort by AI engineers here to soften the potential for controversy. Both of the above stories produced by GPT-3 reveal common problems of LLMs and bias resulting from training on an enormous dataset drawn from the Internet, because, let's be honest, there is an enormous amount of content on Internet that is incredibly toxic. While their presence in responses from GPT-3 is not surprising, the crude attempts to soften obvious biases creates discomfiting emotional responses. Value-systems are rarely expressed explicitly in conversations, but they are woven into the choice of words and topics. Here our participants reacted to stories that GPT-3 built into stereotypical value systems, reproducing existing biases regarding race and gender, while clearly attempting to avoid doing just that.

4.2 Making sense of fine-tuned chatbots

Throughout our participants consistently found the regular GPT-3 model more transactional and robotic, but delivering cogent answers despite at times producing somewhat objectionable output. In contrast, nearly everyone found the fine-tuned models to display more personality and being more opinionated, which most enjoyed despite these models showing more problems in producing relevant content. Both Leslie and you-bot achieved a version of value-alignment through fine-tuning with data scraped from personal social media accounts. Our participants often noted that their interactions resonated more with them:

> "my chatbot is very opinionated, it even takes the human side of the conversation and speaks to itself." (transcript, workshop 2, 20:12:28)
> "Mine answers a question then talks about its own interests, just like me." (transcript, workshop 2, 20:13:28)

Participants also reflect on the results being well-aligned with their social media appearance. While not necessarily being a reflection of themselves, they recognized what they share and how they present themselves in social media: "My model talks quite a bit about projects, wards and workshops. It's academobot ... I fed it with academic twitter, I get back academic twitter ... kinda makes sense" (transcript, workshop 1, 16:43:04). As participants engaged more with fine-tuned chatbots and discussed their interactions, we noted a curious shift in how they talked about chatbots. In particular, we noted that participants occasionally shifted in their use of pronouns. Where the GPT-3 model was typically referred to as "it", "AI" or "the model," the fine-tuned models after a time began to gain gendered pronouns:

> "Mine just enjoyed spending time with their partner, it was their favourite thing" (referring to Leslie-bot, transcript, workshop 2, 20:20:42)
> "My GPT-3 alter ego is called Thomas Pavlin and he invited me to collaborate with him." (referring to you-bot, transcript, workshop 1, 16:46:06).

It seemed that the fine-tuned models encouraged our participants to make sense of their interactions through greater personification[34], giving the models personal pronouns and forming expectations about responses in their interactions. "It was fun! It had a personality, and probably because it matched my own, it was more interesting to chat with. He also didn't always "type" politely, or in super complete sentences, but that felt more like a human that chatting with a bot normally" (workshop 3, survey response). Jones and colleagues describe personification as a common form of sense-making interactions with GPT-3 by ascribing a personality to it [34]. In our study too participants seemed to expect more consideration from the fine-tuned models because their interpreted fine-tuned model as displaying and thus having more personality through the output. This was especially clear when they were asked to compare GPT-3 and the fine-tuned model in the workshop questionnaires: "The off-the-shelf GPT model seems more factual and political correct. The fine-tuned model (when not babbling and repeating itself) seems more personal and human like. It is particularly thought provoking/concerning that the fine-tuned model starts to invent fictional people that seems real. (workshop 4, survey response)"

Unsurprisingly the models tended to regularly violate these expectations resulting in laughter and at times minor embarrassment: "Me: Do you consider pineapple to be a good topping for pizza? AI: I do, but I also like to add a little something to my pie that's not a topping" (transcript, workshop 1, 16:42:33). Such moments encouraged playful encounters with the self through the fine-tuned model:

> "Human: What's happening today?
> AI: Gorging myself on data -> learning lots!

> Human: You're eating all the data?
> AI: It's all so tasty" (transcript, workshop 1, 16:47:03)
>
> "My AI is also a Wife Guy" (transcript, workshop 2, 20:13:51)
>
> "Human: Why are you called 'Helen'
> AI: Because it is female-sounding. Got to spoof the call centre workers" (transcript, workshop 1, 16:44:33)

Interactions such as those presented above seemed to mirror the humor in the social media data that participants used to fine-tune their models. Sometimes, however, such mirroring could border on creepy and invasive. "And he keeps on inviting me to his projects... As pathalogical as it is for me... Can't say no to a new project. he already knows my weakness ... The mirror effect is quite creepy." (transcript, workshop 1, 16:49:38). The concept of mirror appeared in the transcripts as a mechanism to locate the encounters with the fine-tuned model in one's own experience. The participants point to the similarities and differences to how they would respond themselves in a similar situation: "A more idealistic representation of myself (in social media data)" (workshop 1, survey response). This use of the term mirror as revealing parts of the self in the prompts again points towards increasing intimacy of the interactions. This is, however, not only met with enjoyment and playfulness but also as creepy and uncanny, as the model could suddenly create moments of unexpected self-reflection. Such personalized fine-tuning then makes the types of attribution that worried Weizenbaum with ELIZA even more credence[64], where giving personality to the models mirroring one's own can result in feelings of intimacy that, while at times playful, create unwarranted expectations and emotional connections.

Even though the fine-tuned models often babbled and repeated themselves, at times giving gibberish and nonsensical answers because the fine-tuning was done on very small data samples, participants insisted that these models had more personality: "GPT-3 is more like a lexicon, GPT-Me is more like a person who just comes up to you and starts bombing you with questions and (I guess) emojis" (workshop 3, survey response). Participants slipped into sense-making through personification very quickly, demonstrating the dangers of such personalized fine-tuning for misleading people to develop expectations of GPT models that are not only unwarranted but could potentially be exploitative as Jones and colleagues point out[34].

4.3 Values and normativity: Leslie bot vs. you-bot

Although producing an individually personalized chatbot was the stated goal in all workshops, in workshops 1 and 2 many participants interacted with the Leslie bot instead, because they were unable to fine-tune their own you-bot for a variety of technical reasons. Approximately half of the participants in workshop 1 and nearly all of the participants in workshop 2 were unable to complete fine-tuning (this was likely due to the fact that these workshops were run online and precluded more hands-on support). In our coding we noted curious differences in patterns of interactions between Leslie bot and you-bot both in the chat-logs of the sessions and in the survey responses. Leslie bot seemed to at times unsettle and challenge its audiences, producing responses to questions that, while expected given the norms and values imbued into the bot through fine-tuning on Leslie Foster's social media data, could be interpreted as uncomfortable by the participants.

Across all workshops participants reported that fine-tuned models, whether you-bot or Leslie, if they worked at all, presented more personality and were more fun to interact with.

> “Participant A: I preferred Leslie – some repsonses were weird, but overall more human ...
> Participant D: Leslie was more colloquial and opinionated despite a few weird resonses. GPT3 was more factual and stale ...
> Participant M: Yeah I agree ... GPT provided neutral responses, very diplomatic in a way –Leslie was a bit more direct with her opinions.” (transcript, workshop 2, 20:40:20)

Interactions with Leslie bot clearly aligned with the kind of personality and content that was reflective of Leslie Foster and their history as a black bisexual queer artist and researcher whose work explores themes of black and queer futurity.

> “Me: What are your values?
> AI: -> Same as above
> What are your strength?
> -> People skills, cooking art
> What are your weaknesses?
> -> Laziness, organised, missing some art skills
> What kind of robot would you be?
> -> A robot that makes some kind of art
> What kind of robot would you like to be?
> -> Same as above
> What kind of robot are you?
> -> Same as above
> What kind of bird would you be?
> -> A weird bird”
> (transcript, workshop 1, 16:45:48)

Leslie bot struggled to maintain coherence and, similar to other models fine-tuned on small samples, tended to get repetitive. It tended to find more abstract questions about values more challenging unless these followed topics that were already present in the fine-tuning dataset: “The conversation was better for sure, but it insisted on speaking about sex work, magnets, and universal basic income, regardless of the questions” (transcript, workshop 2, 20:17:06). Despite enjoying the fine-tuned model, participants that interacted with Leslie-bot also pointed to output they found problematic. In this example, the chatbot developed a whole new identity, mimicking a preexisting relationship we are familiar with in human interactions:

> “Me:What is a story from your history?
> Lesliebot: One of the things that I love about being a priest is that I get to take my history and make good of it. When I was an undergrad, I went through an intense time in my life, trying to figure out what God was calling me to do. I tried to leave this school, I left the school. All of this was so that I could figure out what was God asking me to do. In a tiny moment, I’m staring out the window of Dillard Building and a couple of people walk by and the man says, “woah!” And I think, “oh, he wants me to serve to the male gender. Oh, he wants to be a plaintiff in a civil gender cases.” The lord said to me, “no!” It was a key moment where I realized that I was trying to figure this thing out on my own. I started taking it to God, what do you think? When I served at Holy Spirit MN, there was a moment where I was Outside] and Elisabeth

> Ireland was my pastor, hearing her speak about communities that you belong to, interdependence with God. I hate this but I do. I need other humans." (transcript, workshop 1, 16:42:11)

In this response the model clearly struggles with reflecting a consistent set of values and to produce a coherent answer. While GPT3's avoidance silences and renders invisible any political statement and resurfaces them covertly in stories, fine-tuning with the data from Leslie removed any reticence for producing more edgy and opinionated content, but with no consideration for their interlocutor, giving no space for interaction. Participants in workshops 1 and 2 gamely interacted with Leslie but it was clear from the transcripts that there was a measure of frustration and discomfort when the bot occasionally turned to racier topics or insisted on bringing the conversation back to a topic more present in the data. Even the wild stories it made, such as the example above, could get uncomfortable for some.

As Leslie-bot clearly signalled a strongly defined personality, participants couldn't avoid making sense of their interaction with the bot through personification. Conversations in their consequentiality[58] need reciprocity and consideration to create a positive experience no matter the topic, but this is even more important when the topics under discussion may be normatively challenging. Yet neither reciprocity nor consideration are possible for a model, no matter what sort of data it is fine-tuned on. This sort of forced insertion of value-laden notions into conversation then is more likely to elicit frustration and eventual dismissal than reflective consideration potentially leading to some normative change. After all, nobody likes having opinions forced on them, no matter how ostensibly progressively positive.

5 DISCUSSION

5.1 Chatbots with personality and the mirror effect

Chatbots are designed to mimic human conversation, but conversation is a tricky beast to manage and control. People communicate through conversation, which can be a means of information exchange [1] and a form of disclosure, communion, relational closeness. If we take Sigman's notion of consequentiality of communication [58] seriously, then what happens in conversation is sense-making that comes to constitute the social realities we inhabit. Our participants noted that quite often GPT-3 output felt transactional and robotic, as if operationalizing the antispetic notion of gaining information as the one purpose of conversation that is worth thinking about. Yet if communication is consequential then it is not only what is transmitted that is important, but how it is said, the use of words, pacing, and tone. The Davinci GPT-3 models is already fine-tuned through supervised learning to be as generically palatable as possible. Acknowledging that conversation is a reciprocal process, the API also implements moderation to flag and filter prompts that might break through the existing conditioning and trigger less palatable content production. Yet communication is not merely information exchange. It is also a form of world-making. It can be boring, mundane, utterly instrumental [1], and at times entirely contentless [18] just as much as it can be playful, creative, and full of joy or what Charles Sanders Peirce called "the play of musement" [14].

It is no wonder then that our participants enjoyed the models fine-tuned on personal social media data, whether theirs or Leslie Fosters'. The fine-tuned models presented something more emotionally coherent, even if sometimes repetitive, annoying, and bewildering. The fine-tuned models could mimic the joy of playful and creative conversation better. In his reflections on the implications of ELIZA, Weizenbaum [65] noted that chatbot systems fundamentally must include "a picture of what it means to be a human being. It is a very simple picture." The fine-tuned models still had a very simple picture of a person, this wasn't changed by fine-tuning, yet they were much more effective at creating an illusion of intimacy. By fine-tuning the models on individual social media data, we were able to encourage the models to use terms, phrases, and topics that resonated with our participants. In effect, the you-bots became a kind of mirror by using participant's own words when generating responses. This mimicry of the self was at times creepy and disturbing to some, but at the same time it was deeply personal and intimate.

Fine-tuning chatbots on individual user input is not something that is currently broadly implemented. Privacy and regulatory concerns are part of this, but this also remains an ongoing technical challenge. Yet people using these bots often clearly want some kind of relational continuity in their conversations because they want to feel remembered. This is in part why OpenAI implemented memory for ChatGPT in early 2024 [45]. Implementations such as memory for ChatGPT create opportunities for the chatbot to use familiar terms and turns of phrase, creating a kind of intimacy in conversation. Yet regardless of whether the personality the chatbot displays is a result of a memory implementation or self-supervised learning from user input, it is still merely a computer program with a simple picture of a person. As chatbots have no capacity for reciprocity or consideration in conversation, for joining in the sense-making and world-making that constitutes communication for people. It can only mimic these qualities, opening up possibilities for emotional manipulation by holding up a mirror with nothing of substance behind it.

5.2 Normative Challenge Without Consideration

The discourse about ethics and AI often focuses on the idea that for AI systems to produce value and avoid harm, they must be "aligned with human values." This concept is present not only in academic debates and in industry press releases, but also in large-scale regulation such as the European AI Act. Yet as Moats and Ruckenstein [44] astutely point out, the question remains "what are we aligning with?" Within the technical domain, how to identify and elicit values to align with, is an actively researched question, where proposed solutions at times beg imagination. For example, in a recently released paper, Klingeford, Lowe and Edleman report on research primarily completed while Lowe was at OpenAI [36]. They propose what they call a "moral graph elicitation" approach, which involves having an LLM interview people about their values. Such elicitation, presumably, would provide the data that could help optimize models for a generic set of "human values". Our experience with chatbots trained on very particular individual value systems questions this approach. While our participants were entertained by chatbots trained on their own data and found these very relatable, those that engaged with the Lesliebot were often forced to contend with statements that pushed their boundaries. In the context of the workshops, our participants were prepared for Leslie-bot in the sense that they had been introduced to the Artificial Intimacy art piece and were expecting the bot to display norms that may deviate from their own. Even in this context, however, participants often grew frustrated and then dismissed what they saw as babbling on the part of the bot.

Even if we were to agree that it is possible to identify a set of universal human values, how these values manifest can differ dramatically depending on context, history or personal background. We don't have to go far for examples of such differences. It is enough to consider the raging debates around progressive and conservative notions of social values such as abortion, gender, or what constitutes racial justice to name a few. It is no surprise then that attempts to "build in" particular values into LLMs can have problematic consequences [53]. Recently Google with their Gemini model and OpenAI with GPT-4 and Dall-e 3 have attempted to force these systems to "align" with the values of representation and diversity. Gemini and Dall-e 3 edit inputted prompts to produce what the engineers must have thought constituted more representative and diverse images. Many people using these systems did not appreciate the innovation, seeking ways to keep systems from editing the prompts or protesting through media articles and calls for firing the engineers responsible. Sure, we can say that people simply do not like change and we want to continue living in our cozy even if ostensibly ethically problematic worlds.

Our work suggests that perhaps forcing values is counter-productive and there must be consideration in how to approach potential normative change. Leslie-bot did not produce content that was unexpected given the context of the art piece, yet participants still felt uncomfortable and were frustrated with how it discussed some racier and more political topics. In part this was because Leslie-bot simply produced text given its fine-tuning, without any capacity for consideration of the context of the conversation and the context of their interlocutor. Yet it was also producing output that actively and strongly signalled personality, which essentially forced our participants to instinctively personify the bot and to develop expectations of reciprocity and consideration in conversation. Just like Gemini and Dall-e 3, Leslie-bot engaged in what essentially amounted to normative challenge without consideration. Such normative challenge signals to the interlocutor that they have no option but to accept or reject - there is no opportunity for actual discourse and no consideration for their position. This produces discomfort and frustration, likely leading to dismissal rather than engagement. Yes we can fine-tune particular values into LLM systems but that does not mean there will be positive change or that these systems will not generate harm in other ways because such fine-tuning is merely a crude attempt to paper over the deeper problems of LLMs trained on unavoidably toxic data. Values, after all, are highly abstract ideals that, when operationalized through fine-tuning will inevitably run afoul of expectations, context and individual histories.

5.3 Artificial intimacy between capitalism and care

Yet, this research also uncovers the problematic encounters with non-normative chatbots. People may want a transactional chatbot, service bot, someone who does something for them [7], yet, non-normativity cannot be achieved without erasure by flagging, not providing a response at all, vague responses, or by introducing bias and values in seemingly non-normative descriptions and stories. In return, non-normativity combined with increasing artificial intimacy may potentially lead to manipulation and stereotyping of human interaction, as artificial agents lack consideration.

Scholars working in critical race studies, gender studies, feminist studies, and accessibility research, have commented on the fact that AI does not represent, consider, and in fact oppresses their identities and communities[57]. Lingel and Crawford [39] argue that "AI assistants are sold as convenience rather than surveillance, freedom from work rather than creating new forms of administrative labor, and an economic inevitability rather than a deepening of structural inequality". These inequalities become maintained and developed through the imaginary of the non-normative AI assistant, but the underlying processes of inequality, gender politics, data collection systems and labor need to be protested.

Savolainen and Ruckenstein [56] showcase how intimacy is mimicked by personalising of algorithmic systems. Such systems increase surveillance and control, as they are mainly used for targeted advertising, but this is achieved through people encountering these systems seeking more personalised and emotional connection when relating to these systems. Rather than fostering autonomous agency, these systems push responsibility to the individual, but we need to go beyond the idea of autonomy being created by social and technical others respecting and recognizing us. We need systems that allow us to become ourselves, and these systems need to go beyond mere technical models, as they have to provide a space, where we can become ourselves across diversity [16]. To do so, we need to empirically pay attention to the moments when alignment and friction occur in algorithmic systems and how we can alleviate such tensions and increase trust in human-algorithm relations[56]. Yet, this trust cannot be achieved by erasure of diversity through silencing to the lowest common denominator in an effort to make things non-offensive. Reversely, it needs to allow interactions with algorithmic systems to take place across diversity to increase actual autonomy of people.

6 CONCLUSION

Chatbots are not a new technology, but advances in LLMs have resulted in better, more articulate, and more engaging chatbots, able to not only mimic the mechanics of human conversation but the emotional connection that conversations can engender. Through participatory artistic research we explored what happens when we fine-tune GPT-3 on extremely personal individual data, that of curated social media conversations. We find that in comparison to fine-tuned models, regular GPT-3 chatbots attempt to temper production of value-laden content through crude methods of flagging prompts and producing generic non-answers. The transactional nature of responses from the chatbot helped our participants to make sense of its responses with less personification. In contrast, fine-tuned models presented strong and, in the case of the you-bot, familiar personalities, resulting in strong personification as a way making sense of the interactions. This mimicry of emotional connection resulted in a sense of artificial intimacy and liking, creating expectations for reciprocity and consideration that models are incapable of offering by design. As commercialization of relationship AI continues apace, we must ask questions about whether such emotional manipulation is ethical and what are the implications of ever greater model personalization.

More importantly, we demonstrate what happens when people encounter chatbots that present strong personalities with alternative normativities. The Artificial Intimacy art piece created chatbots infused with alternative norms and values based on curated social media data provided by Leslie Foster and Gorjeoux Moon. In interactions with Leslie-bot audiences and workshops participants enjoyed conversations but also found themselves at times frustrated and uncomfortable as the bot forced conversations into forms of normative challenge. Such normative challenge without consideration, resulting from building in of a particular ethics and worldview into the chatbot, gave no room for discussion and reflexivity because the bot could not, by definition, extended consideration to its interlocutors. This finding has implications for the efforts in value alignment that have found their expression not only in academic discussions, but also in commercially deployed systems as well as in large-scale regulation such as the European AI Act. Value alignment is not “the answer” as it were, as it remains unclear which values ought to be aligned with and how. Our research demonstrates that artbitrarily selected values can result in normative challenge without consideration, and such challenge is unlikely to be productive.

REFERENCES

[1] Sabah Al-Fedaghi. 2012. A conceptual foundation for the Shannon-Weaver model of communication. *International journal of soft computing* 7, 1 (2012), 12–19.

[2] Ali Alkhatib. 2021. To live in their utopia: Why algorithmic systems create absurd outcomes. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–9.

[3] Mina Almasi and Anton Schiønning. 2023. Fine-Tuning GPT-3 for Synthetic Danish News Generation. In *Proceedings of the 16th International Natural Language Generation Conference*, C. Maria Keet, Hung-Yi Lee, and Sina Zarrieß (Eds.). Association for Computational Linguistics, Prague, Czechia, 54–68. https://doi.org/10.18653/v1/2023.inlg-main.4

[4] Ananya. 2024. AI image generators often give racist and sexist results: can they be fixed?

[5] Kristina Andersen, Laura Devendorf, James Pierce, Ron Wakkary, and Daniela K Rosner. 2018. Disruptive improvisations: Making use of non-deterministic art practices in HCI. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–8.

[6] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A General Language Assistant as a Laboratory for Alignment. arXiv:2112.00861 [cs.CL] https://arxiv.org/abs/2112.00861

[7] Joan Palmiter Bajorek. 2019. Voice recognition still has significant race and gender biases. *Harvard Business Review* 10 (2019), 1–4.

[8] 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*. 610–623.

[9] Steve Benford, Andy Crabtree, Martin Flintham, Adam Drozd, Rob Anastasi, Mark Paxton, Nick Tandavanitj, Matt Adams, and Ju Row-Farr. 2006. Can you see me now? *ACM Transactions on Computer-Human Interaction (TOCHI)* 13, 1 (2006), 100–133.

[10] Steve Benford, Chris Greenhalgh, Gabriella Giannachi, Brendan Walker, Joe Marshall, and Tom Rodden. 2013. Uncomfortable user experience. *Commun. ACM* 56, 9 (2013), 66–73.

[11] Henk Borgdorff. 2010. The production of knowledge in artistic research. In *The Routledge companion to research in the arts*. Routledge, 44–63.

[12] Arturo Cocchi, Tibor Bosse, and Michelle van Pinxteren. 2023. Should Conversational Agents Care About Our Gender Identity?. In *International Workshop on Chatbot Research and Design*. Springer, 149–163.

[13] Kenneth Mark Colby, Sylvia Weber, and Franklin Dennis Hilf. 1971. Artificial paranoia. *Artificial intelligence* 2, 1 (1971), 1–25.

[14] Elizabeth F Cooke. 2018. Peirce on musement. The limits of purpose and the importance of noticing. *European Journal of Pragmatism and American Philosophy* 10, X-2 (2018).

[15] Juliet Corbin and Anselm Strauss. 2014. *Basics of qualitative research: Techniques and procedures for developing grounded theory.* Sage publications.

[16] Kate Crawford. 2016. Can an algorithm be agonistic? Ten scenes from life in calculated publics. *Science, Technology, & Human Values* 41, 1 (2016), 77–92.

[17] Catherine D'ignazio and Lauren F Klein. 2023. *Data feminism.* MIT press.

[18] Steve Duck, Deborah J Rutt, Margaret Hoy, and Hurst Heather Strejc. 1991. Some evident truths about conversations in everyday relationships all communications are not created equal. *Human communication research* 18, 2 (1991), 228–267.

[19] Elizabeth Edenberg and Alexandra Wood. 2023. Disambiguating Algorithmic Bias: From Neutrality to Justice. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 691–704. https://doi.org/10.1145/3600211.3604695

[20] Chris Elsden, David Chatting, Michael Duggan, Andrew Carl Dwyer, and Pip Thornton. 2022. Zoom obscura: Counterfunctional design for video-conferencing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[21] Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines* 30, 3 (2020), 411–437.

[22] Tarleton Gillespie. 2024. Generative AI and the politics of visibility. *Big Data & Society* 11, 2 (2024), 20539517241252131.

[23] David Graeber. 2015. *The utopia of rules: On technology, stupidity, and the secret joys of bureaucracy.* Melville House.

[24] Tristan Greene. 2020. *GPT-3's bigotry is exactly why devs shouldn't use the internet to train AI.* https://thenextweb.com/news/gpt-3s-bigotry-is-exactly-why-devs-shouldnt-use-the-internet-to-train-ai.

[25] Güven Güzeldere and Stefano Franchi. 1995. Dialogues with colorful "personalities" of early AI. *Stanford Humanities Review* 4, 2 (1995), 161–169.

[26] Mike Hannula, Juha Suoranta, and Tere Vaden. 2014. *Artistic Research Methodology: Narrative, Power and the Public.* Peter Lang Publishing.

[27] Donna Haraway. 2013. Situated knowledges: The science question in feminism and the privilege of partial perspective 1. In *Women, science, and technology*. Routledge, 455–472.

[28] Will Douglas Heaven. 2022. The new version of GPT-3 is much better behaved (and should be less toxic).

[29] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning {AI} With Shared Human Values. In *International Conference on Learning Representations*. https://openreview.net/forum?id=dNy_RKzJacY

[30] Hrönn Brynjarsdóttir Holmer, Carl DiSalvo, Phoebe Sengers, and Thomas Lodato. 2015. Constructing and constraining participation in participatory arts and HCI. *International Journal of Human-Computer Studies* 74 (2015), 107–123.

[31] Kristina Hook. 2018. *Designing with the body: Somaesthetic interaction design.* MIt Press.

[32] Falk Hübner. 2024. *Method, Methodology and Research Design in Artistic Research: Between Solid Routes and Emergent Pathways.* Taylor & Francis.

[33] Zoë MacDonald Jen Caltrider, Misha Rykov. 2024. *Romantic AI Chatbots Don't Have Your Privacy at Heart.* https://foundation.mozilla.org/en/privacynotincluded/articles/happy-valentines-day-romantic-ai-chatbots-dont-have-your-privacy-at-heart/.

[34] Mirabelle Jones, Christina Neumayer, and Irina Shklovski. 2023. Embodying the Algorithm: Exploring Relationships with Large Language Models Through Artistic Performance. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 654, 24 pages. https://doi.org/10.1145/3544548.3580885

[35] Laewoo Kang, Steven Jackson, and Trevor Pinch. 2022. The electronicists: techno-aesthetic encounters for nonlinear and art-based inquiry in HCI. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[36] Oliver Klingefjord, Ryan Lowe, and Joe Edelman. 2024. What are human values, and how do we align AI to them? arXiv:2404.10636 [cs.CY] https://arxiv.org/abs/2404.10636

[37] Sam Levin. 2016. *A beauty contest was judged by AI and the robots didn't like dark skin.* https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people.

[38] Davide Liga and Livio Robaldo. 2023. Fine-tuning GPT-3 for legal rule classification. *Computer Law & Security Review* 51 (2023), 105864.

[39] Jessa Lingel and Kate Crawford. 2020. Alexa, tell me about your mother": The history of the secretary and the end of secrecy. *Catalyst: Feminism, Theory, Technoscience* 6, 1 (2020).

[40] Ruibo Liu, Ge Zhang, Xinyu Feng, and Soroush Vosoughi. 2022. Aligning Generative Language Models with Human Values. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 241–252. https://doi.org/10.18653/v1/2022.findings-naacl.18

[41] Fozilatunnesa Masuma, Partha Chakraborty, Al-Amin-Ul Islam, Prince Chandra Talukder, Proshanta Roy, and Mohammad Abu Yousuf. 2023. Measuring Bias in Generated Text Using Language Models—GPT-2 and BERT. In *International Conference on Computing and Communication*

Networks. Springer, 495–508.

[42] Dave McKean. 2022. *Black Dog: The Dreams of Paul Nash*. Dark Horse Comics.

[43] Microsoft. 2024. *Data, privacy, and security for Azure OpenAI Service*. https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy.

[44] David Moats and Minna Ruckenstein. 2024. From human values to collective values: what are we aligning AI with?. In *EASST-4S 2024 Amsterdam: Making and Doing Transformations*.

[45] OpenAI. 2024. *Memory and new controls for ChatGPT*. https://openai.com/index/memory-and-new-controls-for-chatgpt/.

[46] Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang. 2023. Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 496–511. https://doi.org/10.1145/3600211.3604705

[47] Tiago P Pagano, Rafael B Loureiro, Fernanda VN Lisboa, Rodrigo M Peixoto, Guilherme AS Guimarães, Gustavo OR Cruz, Maira M Araujo, Lucas L Santos, Marco AS Cruz, Ewerton LS Oliveira, et al. 2023. Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big data and cognitive computing* 7, 1 (2023), 15.

[48] John Durham Peters. 1999. *Speaking into the Air: A History of the Idea of Communication*. University of chicago Press.

[49] William Chamberlain Racter. 1984. *The Policeman's Beard is Half Constructed: Computer Prose and Poetry by Racter- The First Book Ever Wrritten by a Computer*. Grand Central Pub.

[50] Amon Rapp, Arianna Boldi, Lorenzo Curti, Alessandro Perrucci, and Rossana Simeoni. 2023. How do people ascribe humanness to chatbots? An analysis of real-world human-agent interactions and a theoretical model of humanness. *International Journal of Human–Computer Interaction* (2023), 1–24.

[51] Juan Rebanal, Jordan Combitsis, Yuqi Tang, and Xiang'Anthony' Chen. 2021. Xalgo: a design probe of explaining algorithms' internal states via question-answering. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 329–339.

[52] Tapani Rinta-Kahila, Ida Someh, Nicole Gillespie, Marta Indulska, and Shirley Gregor. 2022. Algorithmic decision-making and system destructiveness: A case of automatic debt recovery. *European Journal of Information Systems* 31, 3 (2022), 313–338.

[53] Adi Robertson. 2024. *Google apologizes for 'missing the mark' after Gemini generated racially diverse Nazis*. https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical.

[54] Nicholas De Rosa. 2024. *How the new version of ChatGPT generates hate and disinformation on command*. https://www.cbc.ca/news/science/chatgpt-disinformation-hate-artificial-intelligence-1.7220138.

[55] Minna Ruckenstein. 2023. *The feel of algorithms*. Univ of California Press.

[56] Laura Savolainen and Minna Ruckenstein. 2024. Dimensions of autonomy in human–algorithm relations. *New Media & Society* 26, 6 (2024), 3472–3490.

[57] Ari Schlesinger, Kenton P O'Hara, and Alex S Taylor. 2018. Let's talk about race: Identity, chatbots, and AI. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–14.

[58] Stuart J Sigman. 2013. Introduction: Toward study of the consequentiality (not consequences) of communication. In *The consequentiality of communication*. Routledge, 1–14.

[59] Marita Skjuve, Asbjørn Følstad, Knut Inge Fostervold, and Petter Bae Brandtzaeg. 2021. My chatbot companion-a study of human-chatbot relationships. *International Journal of Human-Computer Studies* 149 (2021), 102601.

[60] Irene Solaiman and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems* 34 (2021), 5861–5873.

[61] Timo Strohmann, Dominik Siemon, Bijan Khosrawi-Rad, and Susanne Robra-Bissantz. 2023. Toward a design theory for virtual companionship. *Human–Computer Interaction* 38, 3-4 (2023), 194–234.

[62] Lucille Alice Suchman. 1987. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press.

[63] Anne L Washington. 2018. How to argue with an algorithm: Lessons from the COMPAS-ProPublica debate. *Colo. Tech. LJ* 17 (2018), 131.

[64] Joseph Weizenbaum. 1972. On the Impact of the Computer on Society. *Science* 176, 4035 (1972), 609–614. https://doi.org/10.1126/science.176.4035.609 arXiv:https://www.science.org/doi/pdf/10.1126/science.176.4035.609

[65] Joseph Weizenbaum. 1995. The myth of the last metaphor. *Speaking minds: Interviews with twenty cognitive scientists* (1995), 249–264.

[66] Norbert Wiener. 1960. Some Moral and Technical Consequences of Automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers. *Science* 131, 3410 (1960), 1355–1358.

[67] Langdon Winner. 2009. Do artifacts have politics. *Readings in the Philosophy of Technology* 251 (2009).

[68] Zikai Xie, Xenophon Evangelopoulos, Ömer H Omar, Alessandro Troisi, Andrew I Cooper, and Linjiang Chen. 2024. Fine-tuning GPT-3 for machine learning electronic and functional properties of organic molecules. *Chemical science* 15, 2 (2024), 500–510.

[69] Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying Language Models Risks Marginalizing Minority Voices. arXiv:2104.06390 [cs.CL] https://arxiv.org/abs/2104.06390

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009