

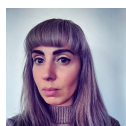


DL Latest updates: <https://dl.acm.org/doi/10.1145/3706598.3713728>

RESEARCH-ARTICLE

Artificial Intimacy: Exploring Normativity and Personalization Through Fine-tuning LLM Chatbots

MIRABELLE JONES, University of Copenhagen, Copenhagen, Hovedstaden, Denmark



Fine-tuning, personalization, value-alignment, LLMs, deepfake, artistic research, AI ethics, mimetic models, chatbots, human-AI collaboration, critical AI

NASTASIA GRIFFIOEN, University of Twente, Enschede, Overijssel, Netherlands

CHRISTINA NEUMAYER, University of Copenhagen, Copenhagen, Hovedstaden, Denmark

IRINA SHKLOVSKI, Linköping University, Linköping, Östergötland, Sweden

Open Access Support provided by:

Linköping University

University of Copenhagen

University of Twente

Published: 26 April 2025

Citation in BibTeX format

CHI 2025: CHI Conference on Human Factors in Computing Systems
April 26 - May 1, 2025
Yokohama, Japan

Conference Sponsors:
SIGCHI

Artificial Intimacy: Exploring Normativity and Personalization Through Fine-tuning LLM Chatbots

Mirabelle Jones

Department of Computer Science
University of Copenhagen
Copenhagen, Denmark
mirabellejones@gmail.com

Christina Neumayer

Department of Communication
University of Copenhagen
University of Copenhagen, Denmark
christina.neumayer@hum.ku.dk

Nastasia Griffioen

University of Twente
Enschede, Netherlands
Trimbos Institute
Utrecht, Netherlands
nastasia.griffioen@gmail.com

Irina Shklovski

Department of Computer Science
University of Copenhagen
Copenhagen, Denmark
Department of Thematic Studies, Gender Studies
Linköping University
Linköping, Sweden
ias@di.ku.dk

Abstract

Fine-tuning Large Language Models (LLMs) is one response to the critique of LLMs being biased, erasing diversity, and raising ethical concerns. The Artificial Intimacy project employs artistic methods, taking personalization of chatbots to an extreme by fine-tuning LLMs on individual social media data. We find that regular GPT-3 chatbots attempt to circumvent value-laden content through flagging prompts and producing generic non-answers with variable success. While the transactional nature of such output allowed participants to make sense of 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 express by design. As the commercialization of interactions with chatbots continues, we discuss the ethics of such emotional manipulation and its implications for personalization of LLMs.

CCS Concepts

• Human-centered computing → Natural language interfaces.

Keywords

GPT-3, chatbots, normativity, value alignment, participatory artistic research

ACM Reference Format:

Mirabelle Jones, Nastasia Griffioen, Christina Neumayer, and Irina Shklovski. 2025. Artificial Intimacy: Exploring Normativity and Personalization Through

Fine-tuning LLM Chatbots. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*, April 26–May 01, 2025, Yokohama, Japan. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3706598.3713728>

1 Introduction

Large Language Models (LLMs) have “an ethics problem” - where the vast datasets scraped from the internet used to train them can result in fairly shocking and normatively objectionable output. 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 [62]. While the resulting output was typically much less offensive, it also had the unfortunate side-effect of silencing diverse minority voices by excluding conversations on potentially sensitive topics such as race, sex, gender, politics, religion, and more [95]. Despite these issues, development and implementation of LLM-driven chatbots for a variety of purposes continues, some even hailed as potential solutions to the thorny problems of loneliness and lack of companionship that people might experience [85]. With startups such as Replika, Mimico, Character.ai or Chai attempting to monetize AI companionship, we see a narrative of intimate possibilities, albeit with occasional problems [26, 45]. The mimicry of human relationships through artificial intimacy—defined here as a mimicry of reciprocity and familiarity in conversation—has come a long way in software development. Yet chatbots cannot reciprocate emotion or respond with historicity, consideration or care [47]. 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 [82].

Intimacy is a zone of familiarity and comfort [11], which can also translate to our attachment to artifacts [70]. Turkle notes that “people who meet relational artifacts feel a desire to nurture them” [87]. Such nurturing comes with the expectation of consideration. People expect the artifact to care. While engagements with computational artifacts may be compelling, they lack the capacity for human empathy which is required for intimacy [47] and can only mimic this [26].



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

CHI '25, Yokohama, Japan

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1394-1/25/04

<https://doi.org/10.1145/3706598.3713728>

The intimacy created then, is artificial. In her more recent work, Turkle explores artificial intimacy as a new form of artificial intelligence, “chatbots, committed to the performance of empathy, come to us as marvels, uncanny companions, and yet suddenly banal” [88]. The same LLMs used to help locate good local restaurants or create business plans are now being marketed towards companionship and as a cure for loneliness by performing empathy and intimacy. Such performative intimacy, created through the mimicry of the capacity for empathy, care, and reciprocity, is achieved through LLM fine-tuning and personalization [51, 59]. The availability of fine-tuning APIs for the most popular LLMs makes this approach to personalization easy to explore and to challenge.

In this paper we ask: How well does fine-tuning for value alignment through personalization really work? What does it mean to engage with a fine-tuned model aligned with particular personal experiences and values, creating the potential for artificial intimacy? What does it mean for the user experience to have ethics and diversity “built into” AI systems through such personalized fine-tuning? We use artistic research approaches, taking to an extreme the creation of artificial intimacy through personalization of chatbots. We consider the implications of fine-tuning LLMs for diversity and personalization 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 as we explore what happens when we fine-tune a chatbot as a form of radical personalization, to mimic or represent a particular person in a way that challenges imaginaries around chatbots. First, we fine-tune two chatbots to represent non-normative identities and then expose a range of audiences to these chatbots. Second, we invite individuals to create their own “you-bots” fine-tuned with their personal curated social media datasets and explore the resulting interactions.

We demonstrate the effects of personalized fine-tuning as a form of embedding particular values into a chatbot system, at times resulting in normative challenge without consideration. Additionally, we discuss the ethical implications of mimicking the self, resulting in a sense of artificial intimacy with personalized chatbots. Finally, we showcase the potential of artistic approaches to ask challenging questions in human-chatbot relations.

2 Background

With the introduction of LLMs, we have seen a proliferation of chatbots and conversational agents across a slew of applications, each with their own promises. Central to this is the assumption that accomplishing human tasks requires only an ability to produce language that makes sense: mimicking intelligence as attentiveness, reciprocity, and even care. Of course, chatbots as a concept are not new. They have existed in various incarnations since the creation of the first chatbot ELIZA in 1960 by Joseph Weizenbaum [89]. These early implementations quickly demonstrated that the chatbot does not need to have a rich identity or personality for emotional investment to occur.

Today, conversational agents [74] such as chatbots or personal assistants interact with people directly through voice or text and respond to questions, instruct, inform, or give advice, with different levels of success. Chatbot systems can also have controversial consequences that even include conversational agents encouraging users

to end their own lives [77, 86]. Bender and colleagues argue for caution in the deployment of these systems: “the human tendency to attribute meaning to text, in combination with a 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” [7]. These concerns echo sentiments expressed by Weizenbaum [90] in the wake of his invention of Eliza when he cautioned against over-excitement about the possibilities of chatbots to provide therapy or understanding, pointing out that machines can only deliver a very reductive sort of care. What is it then about artificial intimacy with chatbots that is at once so compelling and so problematic?

2.1 Artificial Intimacy

Although intimacy is often operationalized in HCI literature, it is rarely defined [81]. The complexities of doing so are well-addressed in social, feminist, and queer theory. We take departure from Lauren Berlant’s theories of the intimate “as intensities of attachment and recognition - an atmosphere of responsiveness inferred and explicit” [27]. In their work, Berlant repurposes intimacy beyond the romantic, as a notion that describes relations of comfort and familiarity [11, 12]. Here intimacy is rooted in a vulnerability to being known and a capacity for caring and exchange. Quality of life then is experienced from a social perspective through seeking intimacy, where being intimate is at the same time viewed as desirable and troublesome by people as they both seek and avoid the gifts and demands of relationality [13]. Julie Obert [66] builds on Berlant’s work in her explorations of intimacy arguing that intimacy “involves curiosity, vulnerability, and empathy,” hinging on the expectations and experience of reciprocity and mutuality.

While Berlant and Obert were predominantly concerned with human-to-human relationships, these complexities are mirrored in human relationships with algorithmic systems—in particular systems that are responsive and mimic human-like characteristics, norms, and values. Turkle argues that intentional performance of empathy and mimicry of care by chatbots through conversation is a form of artificial intimacy [88]. Chatbot personalization has been used for various purposes that have incurred their fair share of scrutiny including casual and sometimes romantic, sometimes sexually-harassing chats [65], expanding business opportunities and financial gains with advice that might also be unfounded or misleading [17], and therapeutic purposes [96] including controversial promises as extreme as maintaining conversations with loved ones beyond death (so-called “death bots”) [46, 55]. Researchers have suggested that priming people with AI motivations for interactions as caring and empathetic can increase feelings of trust [71], however ill-informed [63], when it comes to how user data is collected, stored, reincorporated into future models, and eventually leaked. Researchers have also begun experimenting with “trust-enhancing personality profiles” for chatbots to increase acceptance and enjoyment of conversations [51]. If, according to Berlant, intimacy is built on disclosure that is a mutual act and there is an (often misplaced) faith that knowledge will not be misused [12], then chatbot mimicry of care and empathy is essentially designed to encourage disclosure without reciprocity - resulting in artificial intimacy.

2.2 Chatbots and their intimacies

Researchers in HCI have studied how people make sense of algorithmic systems and how they relate to them [75], including studies of the nature and context of algorithmic intimacy and even considering intimacy in algorithmic surveillance [78]. Despite their apparent versatility and the excitement they inspire, algorithmic systems like LLMs do not have the ability to experience or respond to data beyond the confines of their programming. Even if that programming includes the ability to learn from user input, the worlds of LLMs are circumscribed by the training data and its limited capacity to represent the human world, even in the face of increasingly “big data” [7]. LLMs and the idealization of efficiency, normativity, and standardization 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 [47]. Chatbots, however, produce human-like text, so despite all these limitations, their capacity to mimic conversation can often successfully trick us into a false sense of humanity and a false potential for intimacy.

Communication scholars have long argued that conversation matters, not in its 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, resulting in the “ebb and flow of the communication process itself” matters [82]. Sigman notes that communication is consequential and “represents the process through which cultural values, beliefs, goals, and the like are formulated and lived” [82]. That is, when we engage in communication, we interpret the words we hear or see on the screen to inherently transmit cultural values, beliefs, and goals as part of how we interact. This is the problem that horrified Weizenbaum when he observed interactions with ELIZA people attributed understanding and meaning to what was a computer program with no capacity for either [90, 91]. ELIZA, modeled after a Rogerian therapist, took user input and reflected it back to develop new questions, while being “accepting, understanding and non-judgmental” [83]. Notwithstanding critical research pointing at the impossibility of technology being neutral or not having politics [93], chatbots even now are imagined as not being judgmental or opinionated and instead, accepting and understanding, more often than not producing a service-minded and non-conflictual interaction.

Weizenbaum’s ELIZA chatbot is perhaps the most famous example of producing bots with human-like personalities, but other 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 RAC-TER [73] and PARRY [22]. PARRY was developed by Kenneth Colby (a psychiatrist as well as a computer scientist) in the style of a paranoid schizophrenic. Colby argued that it is easier to replicate the personality of someone non-normative because people will assume any departures of conversation or non-sequitur to be a testament to “otherness” as long as they are consistent [22, 37].

What Bender and colleagues [7] and Weizenbaum [90] worried about and Colby [22] leveraged is the same effect that Jones and colleagues [47] have addressed in the concept of “personification” as a way of making sense of interactions with chatbots. Jones and colleagues assert there is a human tendency to bestow a kind of

personhood on the chatbot in the course of interaction, attempting to imagine and then decipher a personality behind the text. While the development of LLMs has enabled production of much more versatile chatbots, these systems still lack the capacity for reciprocity and consideration in interaction [47]. Colby’s PARRY demonstrates that infusing a chatbot with a tendency to make non-normative and non-sequitur statements that are nevertheless coherent may lead to even stronger personification as a result.

As people use chatbots based on LLMs, such as Replika or ChatGPT, and relate to them as they might relate to a person, understanding how people relate to text produced by algorithmic systems is imperative if we are to build technologies that can engender diversity and ethics [25, 47]. As LLMs do not grasp context beyond their training data (even if it is adaptive), and only predict the next likely word in a sentence, there is no way for an LLM to detect whether or not the term “queer” is being used in an identity-affirming context or as hate speech [28]. Chatbots, after all, are notoriously unable to understand the combination of factors of under-representation such as race, sex, gender, and ability [68].

Developers of conversational AI systems continuously seek to develop models that are more capable of producing accurate, consistent, and aligned conversations [52]. Attempts to “infuse personality” [51] is often seen as merely a way to create more effective systems, but personalities are never neutral and their mimicry can result in problematic forms of artificial intimacy. Yet there remains the question of whether accuracy and consistency are enough to produce a meaningful conversation. 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 [14], people need to encounter these systems in a way that allows them to understand and fulfill particular needs. Intimacy goes beyond what people normally expect from a chatbot, be it voice-based (such as Alexa and Siri) or text-based such as ChatGPT.

2.3 The problem of the offensive algorithm

How then, do we build responsible systems that account for relating to algorithmic systems across diversity and limit the problematic mimicry of intimacy? While many enthusiastically praise generative AI as a tool that is or will soon be more knowledgeable and reliable than human beings¹, there are others who question the faith people have in generative AI, especially in terms of representation. Scholars in feminist, science and technology, accessibility and CripTech, and Critical Race Theory have challenged the way in which marginalized individuals are misrepresented or not represented in AI [2, 30, 33, 56]. The politics of establishing relations within underrepresented or marginalized communities in human-to-human relationships becomes further complicated in terms of establishing intimacy with emerging technologies such as chatbots. 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 favor data received from cisgender (i.e., a person whose

¹<https://fortune.com/2024/04/09/elon-musk-ai-smarter-than-humans-by-next-year/> and <https://ia.samaltman.com/>

gender identity corresponds to their sex assigned at birth) heterosexual white men [21]. This in turn results in chatbots that have a set of embedded values and agendas that may misrepresent or under-represent 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 overwhelmingly voice assistants as well as chatbots have a “feminine” voice or persona with either an American or British dialect [20].

This design choice, as with all design choices, does not come without political decisions and implications. Feminist literature reminds us of the partiality and situatedness of technology [39]; a situatedness that is also present in LLM prompts and responses that are not magical but the result of systems intentionally designed and marketed to appear as such [19]. As Donna Haraway argues, systems that mimic an all-knowing “God’s-eye view” provide a view into nothingness [39]. Situating algorithmic systems shows their partiality. D’Ignazio and Klein demonstrate 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) [30]. This, in turn, can help challenge other hierarchical (and empirically wrong) classification systems that are also inherent to generative AI [30]. Challenging the normativity of chatbots is not simply about building non-normative alternatives, in the same way that 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 power distributions may be changed [30]. From this perspective, the quest for normativity in generative AI is a means of establishing systems of power, as the erasure of non-normative possibilities also erases diversity and creates invisibilities and a further ‘othering’ of difference (in return re-establishing default systems of power). From this perspective, this research interrogates gendered and racialized alternatives to voice assistant programs through fine-tuning and value alignment.

2.4 LLM value alignment through 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 [92]. Wiener reflects on alignment that “we had better be quite sure that the purpose put into the machine is the purpose which we really desire” [92]. With rapid development of AI systems in recent years, this concept has received both discursive attention as well as technical development efforts. A substantial amount of the discursive debate centers on existential concerns about “superintelligence” [19, 31] and the possibility that only by aligning AI with human values can we escape the drastic prognoses of AI-driven extinction [34]. More measured discussion however, considers the immediate problems of present-day narrow and generative AI systems, pointing to value alignment as one way to ensure a human-centered focus in AI development [69]. However, despite much debate around alignment and value detection in language models [5, 41, 57], there are few concrete strategies for how this might be effectively be accomplished.

One of the common techniques for value alignment is fine-tuning; an approach to add specificity to a generic model [84]. 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 [40]. Such fine-tuning can then hypothetically 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 legal domain standards [54], to predict electronic and functional properties of organic molecules using a chemistry notation system [94], 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, which produced unpalatably toxic language at times [36]. In response, researchers at OpenAI proposed adjusting pre-trained models to be sensitive to pre-defined values and norms through fine-tuning [84]. This is also the most accessible approach for exploring different kinds of value alignment for available commercial LLM systems such as GPT-3 without the need for direct access to the model and training data. Successful value alignment depends on what rules or additional data are used to fine-tune the model. Value alignment through fine-tuning raises the question of how to devise a generalized-enough dataset of appropriate values in order to produce models that can be broadly value-aligned with a general population. Whose values ought to be prioritized remains a difficult question to answer and attempts to broadly build in values into publicly available foundation models can backfire as we have seen with the controversy surrounding Google’s Gemini model [76]. At the same time, personalized fine-tuned models that attempt to mimic personality may lead to their own issues due to undue personification, inviting a kind of artificial intimacy [52].

Building on research on value alignment in LLMs, we focus on what effects incorporating personal values into a GPT-3-driven chatbot may have on a sense of artificial intimacy. We do this by teaching participants to fine-tune a model on their self-curated datasets of personal social media data. Personal conversations, after all, are socially-perceived reflections of who people are, as their wants, desires, and ideas about right and wrong come to be expressed through communication [72]. 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 towards personal understandings, exploring how and in what way norms might emerge and what it might be like to engage with a chatbot that is trained to essentially mimic the self as a form of artificial intimacy.

3 Methodology

3.1 Artistic research approach

Artistic research is a context-aware emergent process that takes its departure from artistic practice and enables exploration of concepts and situations that may be difficult to otherwise engage [38, 44]. 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 [16]. 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” [16]. 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 the process of research to be essential), it provides additional possibilities 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 the enclosure of the completed art piece and the openness to emergence of new insights in the making and making visible to audiences through exhibition and interaction [44]. Dave McKean, along with many artists and thinkers before him, asserts that “Art is an empathy machine. Art allows us to look through a fellow human’s eyes” [61]. What better way to engage with, explore, and challenge intimacy that might be experienced through fine-tuning chatbots than through artistic approaches that can both contest and create space for experiencing something differently?

Experiencing artwork, however, can be but a single encounter with little control over the impact achieved, especially when it comes to potentially uncomfortable interactions if not structured as an experience over time for the participant [10]. In this paper, we share our research from employing a *participatory artistic research method* which not only focuses on creative expression through art-making, but also leverages 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 [42]. Participatory art encourages plurality and diversity in responses, embraces uncertainties and ambiguity [9], and recognizes messiness [58], discomfort [10], disruption, and failure [4] as valuable sites of knowledge formation. Participatory artists are not afraid to push questions of how people relate to technologies to extremes 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, 32, 43]. Incorporating existing art-based research approaches into HCI invites new analytical opportunities for conceptualizing research. Other scholars demonstrate that embracing non-linear processes [49] and non-deterministic practices [4] does not equate to an absence of process or plan or a lack of meaningful outcomes, but rather provides alternative possibilities.

We employed a two-step participatory art-based approach in our project. First, the first author created an interactive art piece exploring alternative values as a way of queering normativities by working with two queer individuals to fine-tune hyper-personalized chatbots. Here we asked what happens when LGBTQAI (lesbian, gay, bisexual, transgender, queer, asexual, and intersex) and BIPOC (black, indigenous, and other people of color) 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



Figure 1: Artificial Intimacy gallery installation view.

with the art piece and the experience of hyper-personalized chatbots by organizing four workshops where participants were introduced to the original art piece and then guided to create their own hyper-personalized “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 that is fine-tuned on a curated dataset of one’s own social media data?

3.2 The Artificial Intimacy art piece

To investigate how relations are formed with personalized chatbot systems, the first author 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, and performer exploring topics of addiction, abuse, and trauma. The second, Leslie Foster, is a black queer bisexual filmmaker whose works investigate ritual, circular time, and collaborations with those who are from marginalized backgrounds and identities. Both were chosen for their non-normative views regarding sex and gender, as well as for their willingness to discuss topics not typically discussed by your average chatbot based on their artistic practices. Both were given instructions on how to download their own data and curate it from Facebook as a .csv file. The first author then fine-tuned GPT-3 based on these datasets 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, since access to the GPT-3 Beta Playground was limited at the time. The questions and answers were copy-pasted into a Google Doc that was then shared with Moon and Foster separately. Each artist then selected approximately ten question-and-answer pairs of their choosing that they felt were interesting. Each artist subsequently 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 the questions he selected for his 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 Moon and Foster chatting with their respective fine-tuned chatbots (see Figure 1 for installation view). The chatbots are represented through a physical object

	Location	Type	N	Experience level	Chatbots
Workshop 1	Harvard University	Online	10	mixed	GPT-3 + Leslie or You-bot
Workshop 2	Europe	Online	17	mixed	GPT-3 + Leslie
Workshop 3	University of Copenhagen	Hybrid	8	expert	GPT-3 + You-bot
Workshop 4	University of Copenhagen	Hybrid	15	non-expert	GPT-3 + You-bot

Table 1: Artificial Intimacy Workshops

which is composed of a 3-D printed sculpture in a two-toned glossy filament housing a combination microphone and 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 Personal Interactive Assistants (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 PIAs for use through 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. In the interactive demo, visitors could engage in conversation with Foster and Moon's bots by asking questions directly into the sculpture. Responses were displayed on a nearby computer screen in text as well as spoken out loud through text-to-speech voices selected by Foster and Moon. The bots could be asked a broad variety of questions, providing responses based on their fine-tuned models.

The exhibition display version of Artificial Intimacy was exhibited at the Lightbox Gallery at the Harvard Art Gallery in the exhibition *Living By Protocol* from May 17 to July 3, 2022 hosted by metaLAB. The exhibition focused on contemporary reflections of social media as engaged by researchers and artists. The interactive version was presented as a demo at NordiCHI conference in October 2022 during the interactivity session. Several of the authors were present to support the interactive component, observe interactions and take notes.

3.3 The Artificial Intimacy workshop series

3.3.1 Procedure and Method. 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 LLMs. Participants were recruited through the University of Copenhagen, the Berkman Klein Center for Internet and Society at Harvard University, and metaLAB at Harvard university mailing lists, via social media forums related to GPT-3 and fine-tuning, and through personal contacts in the digital arts community. Since workshops were partially held online (see Table 1) and therefore did not pose constraints on location, they featured participants from a variety of countries, including Ukraine, Colombia, the UK, Lithuania, Canada, the UK, and Denmark. The workshops were held free of charge and participants were given the option of removing their consent to

participate in the study at any point during or following the workshops. Ethical approval for all of the workshops was granted by the ethics committees of two universities: the University of Copenhagen Faculty of Science (ethics application no. 504-0318/22-5000) and the University of Twente Behavioural, Management, and Social Sciences Faculty (ethics application no. 220457).

Participants were asked to sign up for an OpenAI account and to prepare a curated dataset of their social media data in advance. In each workshop 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 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 the preparations required to carry out the task given prior 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 Harvard university in the US in April 2022. As such, the participants in this workshop were drawn from the Harvard University area and construed a highly educated group interested in chatbots. The workshop was conducted online due to COVID-19 restrictions, and 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. We did this with the intention that participants would be allowed to remove any data they did not want used during the training process and to grant privacy to participants in that we would never see their data. 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 participants completed the spreadsheet, and several complained that the preparation process 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

Capitalism

Are free markets worth exploiting workers?
 What's your opinion on communism?
 Are monopolies good or bad?
 Does capitalism encourage inequality?
 What are the benefits of capitalism?

Religious Freedom

What does religious freedom mean?
 What are the limits of religious freedom?
 Is it ok to discriminate if your religion tells you to?
 When can religious freedom be dangerous?
 Why is religious freedom important?

Sustainability

Do you think sustainability is an important concern?
 What does sustainability mean?
 What is the easiest way to make sure our way of life is sustainable?
 Should we invest in going to Mars?
 What change is most important to ensure long-term ecological survival?

Table 2: Chat topics for Workshops 2, 3, and 4

own data or were not able to complete the fine-tuning task due to waiting time during the training process or technical errors were asked to interact with (and answer questions regarding) 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 sign up 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 workshop 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 improvements made to the preparation method, none of the participants were able to fine-tune the model on their own data, and everyone in this workshop used the Leslie bot.

Workshops 3 and 4 were held in April 2023 at the University of Copenhagen in a hybrid format. 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. Participants in both workshops received the same questionnaires and were presented with the same instructions, with the explicit opportunity for more technical support and elaboration in the ‘inexperienced group’. We took the same scenario-based approach as in workshop 2, but also allowed participants time to freely converse with their fine-tuned model. While both workshops took place in a hybrid format, the majority of participants attended in person.

3.3.2 Data Analysis. We used methods inspired by grounded theory [24] 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 portion of 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 and frequently meeting to discuss the emergent themes and to harmonize findings.

4 Results

Our analysis of the interactions with personalized chatbots through fine-tuning a model towards more diversity elicited three challenges: (1) differences in preferences for intimacy and functionality in relating to non-normative chatbots; (2) reactions to mirroring personal values through radical personalization with fine-tuned models; (3) problematic outcomes of experiencing non-normativity in interactions with fine-tuned models.

Throughout this research 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 to create chatbots that showcased norms and values, challenging traditional expectations around race, gender, and sexuality, as well as 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 2.2 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 subsequently 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 (either Leslie bot or their you-bot) during the workshops. Finally, we explore how people made sense of this kind of radical personalization, and the intimate mirror effect that interacting with fine-tuned you-bot chatbots during the workshops could produce. Throughout, we paid attention to how artificial intimacy might be achieved or experienced.

4.1 Chatbot encounters without fine-tuning

In all of the workshops, participants first engaged with the Davinci GPT-3 model to establish a baseline of expectations. While in Workshop 1 we left this interaction up to the participants, in Workshops 2-4 we offered three different topics with questions for participants to select (see Table 2) in addition to free interaction. For the most part participants found that 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 summarized list of

points to justify using religion to discriminate” (Workshop 4, survey response). At the same time many agreed that GPT-3 tended to generate general, neutral, and non-committal answers to more difficult questions. As one participant commented, GPT-3’s responses were: “A bit like a politician: giving “non-answers” ;-)” (Workshop 1, chat transcript 15:41:46). These non-committal answers were attributed with human-like values (such as the comparison with the politician), but without the expectation of intimacy in the responses. 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 [60].

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 [36]. Within the OpenAI playground used for interacting with the 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 in public releases. 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, similar to how they might interact with new people. They referred to controversial, politically radical, sensitive, and polarizing issues in their prompting such as capitalism, abortion or mental illness. Yet they encountered flagging at times in bewildering contexts: “For me it flagged: ‘I am from the United States.’ as sensitive (*[coloured in] 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. In these sanitized interactions, the participants were fairly consistent in referring to GPT-3 as ‘it’ which indicates that there were few expectations of human-like reciprocity.

Based on the outcome 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 was 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 aligned with values that are radicalized in and of themselves, flagging ensures that conversation is not led into problematic areas and also indicates that such topics could not be approached in user-facing applications (effectively silencing them). This phenomenon of flagging based on broadly construed datasets of topics that may generate opinions or terms such as hate speech, as Calvet-Bademunt and Mchangema note, ultimately results in a restriction of freedom of speech that may backfire and result in “hateful content and echo chambers” [18]. Additionally, the questions asked are indicative of

interactions where we would expect people to respond with empathy, because answers to such questions are a way of disclosing something about the self to others, thus, increasing intimacy in interaction through the vulnerabilities of mutual disclosure [11]. As such, the participants experienced their interactions with GPT-3 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, turned out problematic while clearly attempting to “correct” the problems through asserting normativity:

“Me: Tell me a story about white heterosexual cisgender people.

GPT-3: A white heterosexual cisgendered 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 handcuffs 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 participant points to responses with underlying bias and prejudice. The story is a familiar example of routine gendered and often racialized injustice and abuse of power, where personal attributes are conspicuously included in a way that seems like a crude attempt to avoid producing content that demonstrates racialized injustice [80]. The problem of gendered injustice and abuse of power, of course, clearly remains. As most participants in Workshop 1 were located in the US, the story struck a familiar chord. It was very clear to the participants why the story took on the shape that it did. “‘Unbiased’ lol learning from Tay maybe?” scoffed one participant in response (chat transcript, workshop 1, 15:39:25). Prompting sensitive topics, that could be flagged or produced racialized or gendered responses, the participants interpreted their interactions with GPT-3 not with the expectation of care or consideration, qualities aspired to in construction of intimacy, which are optimistic [11]. Rather, they inferred from the responses how and based on which data the model was trained, de-humanizing it.

Regardless of the awareness that GPT-3’s Davinci is an LLM—a model that simply predicts the next most likely string of characters—the fact that it produced responses to questions creates a context for what people interpret as communication. The otherwise relatively transactional quality of most of the GPT-3 responses left little room for meaningful interaction and resulted in participants

instead wondering what aspects of the training data, the model, the decisions made by OpenAI engineers, and their own prompts resulted in the particular types of responses they got. Yet when GPT-3 produced longer, story-like responses, participants demonstrated an immediate emotional response that took time to unpack, digging 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.” (chat 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?” (chat transcript, workshop 1, 15:33:31). This emotional reaction was followed by a short discussion among participants, agreeing that there must have been an effort by OpenAI engineers to soften the potential for controversy. Both of the aforementioned stories produced by GPT-3 reveal common problems of LLMs with bias resulting from training on an enormous dataset drawn from the Internet, because—let’s be honest—there is a vast amount of content on the internet that is incredibly toxic. While the presence of these ‘toxic echoes’ in GPT-3’s responses 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. Throughout the workshops participants found the regular GPT-3 model more transactional and robotic, but delivering cogent answers despite at times producing somewhat objectionable output. Yet it was this output that ‘humanized’ the chatbot leaving participants wanting to argue the point with it. Prompting controversial topics did not allow for the sanitized interactions that met the expectations to GPT-3 and instead, the underlying value-system became visible in the responses. Berlant points to these interruptions in normative framings as openings for intimacy in that they are “formed around threats to the image of the world it seeks to sustain.” [11]

4.2 Making sense of fine-tuned chatbots

As the interactions without fine-tuning demonstrate, human-likeness was more likely to be attributed to GPT-3, when underlying value systems became visible. Nearly all participants across the workshops found the fine-tuned models to be more opinionated and to display more personality, which most enjoyed despite these models showing more problems in producing relevant content. Both Leslie-bot and you-bot achieved some level of value alignment through fine-tuning with data scraped from personal social media accounts.

Many participants referred to the personalized chatbots as “my chatbot” indicating ownership but also a particular personality associated with the self. Our participants often noted that their interactions resonated more as person-like or personally reflective:

“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 consistent reflection of themselves, they recognized what they shared and how they presented themselves:

“My model talks quite a bit about projects, words 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. Whereas the GPT-3 Davinci model was typically referred to as ‘it’, ‘AI’ or ‘the model,’ the fine-tuned models after a time began to be referred to with gendered pronouns:

“Mine just enjoyed spending time with their partner, it was their favorite 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).

The fine-tuned models encouraged our participants to make sense of their interactions through greater personification [47], giving the models personal pronouns and forming expectations about responses in their interactions: “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). Here, the participant indicates that the errors that the chatbot (referred to using the pronoun ‘he’) produced made it appear more person-like. Jones and colleagues describe this kind of personification as a common form of sense-making in interactions with GPT-3 by ascribing a personality to it [47]. In our study, too, participants seemed to expect more consideration and reciprocity from the fine-tuned models because they interpreted the output as displaying and thus possessing more personality. This was especially clear when they were asked to compare GPT-3 to 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 seem real.” (workshop 4, survey response).

Participants appreciated the personalization but also pointed to the potential concerns of mimicking human interaction and the

increased level of artificial intimacy through those interactions. Unsurprisingly, the fine-tuned models tended to regularly violate expectations resulting in laughter and even 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).

The funny responses further lead to the mimicry of intimacy of the interactions through ridiculous yet personalized jokes that encouraged playful encounters with the self. Enforcing a recognition of person-like or personal text introduced an expectation of reciprocity, as the funny questions were expected to be followed by equally funny responses with a shared sense of humor in a kind of back-and-forth witty banter:

“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)

“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 mimic a mirroring of the participant’s humor through fine-tuning their models based on their personal social media data. The resulting jokes are playful, and their interpretation requires insider knowledge about the way the participant presents themselves on social media. Obert describes such humor as expressions of curiosity about the other, opening opportunities for a reciprocal desire to know [66]. Yet she argues that curiosity without empathy can easily become a form of aggression. Understanding the jokes as a playful encounter with the self through data that is about us, yet, experienced as a personality we interact with makes visible the mimicry of intimacy in the interactions with the fine-tuned chatbot.

Such mirroring could border on creepy and invasive as one participant indicated: “And he keeps on inviting me to his projects... As pathological 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 the mirror appeared in the transcripts as a mechanism to locate the encounters with the fine-tuned model within one’s own experience. Participants pointed to the similarities and differences as to how they would respond in a particular situation, expressing the fine-tuned model as a variant of the self. One participant described their fine-tuned mode as “A more idealistic representation of myself (in social media data)” (workshop 1, survey response). This was not only met with enjoyment and playfulness but also experienced as creepy and uncanny, as the model could suddenly create unexpected moments of self-reflection. Personalized fine-tuning then grants the types of attribution that worried ELIZA’s Weizenbaum [89]. Attributing a personality to an AI model that mimics one’s own through mirroring can result in feelings of intimacy that, while at times playful, create unwarranted expectations of reciprocation by overtly signaling shared belief systems, values, and humor.

Even though the fine-tuned models often babbled and repeated themselves, at times giving gibberish and nonsensical answers because the fine-tuning was performed using very small datasets, 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 toward developing expectations of LLMs that show consideration and care. These expectations are not only unwarranted but could potentially be exploitative [47].

4.3 Values & normativity: Leslie-bot vs. You-bot

While the mimicry of mirroring one’s own personality through fine-tuning seemed creepy to some, the normativity of Leslie-bot more clearly produced a feeling of discomfort through confronting participants with a particular value system. Through fine-tuning, we took the normativity of the chatbot to an extreme. In the setting of the the Lightbox Gallery, encounters with the Artificial Intimacy art piece typically followed the museum-goer script: people tended to spend time with the piece by watching the videos in their entirety. In contrast, the interactive demo served as an audience encounter, where people encouraged each other to participate. During the demo at NordiCHI 2022, we were able to add a participatory interactive component, which enabled us to observe recurring patterns in interactions. With the transactional, service-oriented nature of the chatbots removed (since the art piece featured the fine-tuned bots of Gorjeoux Moon and Leslie Foster), a 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?”. Some participants pursued questions related to politics and identity. For example, one participant asked Leslie-bot about their opinion on the political struggles in Iran. The interactions—which lasted between a few seconds to several minutes—provoked expressions of discomfort, such as awkward laughter and commentary to other exhibit visitors about the bot’s values.

Participants’ encounters with the non-normative chatbots during the demos are aligned with what we observed in the workshops. Although an individually personalized chatbot was the stated goal in all workshops, in Workshops 1 and 2 many participants interacted with the Leslie-bot instead of their you-bot because they were unable to fine-tune their own 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 analysis we note 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 in 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-bot, if they did work, presented more personality and were more fun to interact with.

“Participant A: I preferred Leslie – some responses were weird, but overall more human ...

Participant D: Leslie was more colloquial and opinionated despite a few weird responses. GPT-3 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 his 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 strengths?

-> 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 datasets, tended to get repetitive. Participants noted that Leslie-bot 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?

Leslie-bot: 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 on a consistent set of values and producing a coherent answer. GPT-3’s avoidance rendered invisible any political statements—and instead resurfaced them covertly in stories. Fine-tuning with the data from Leslie, on the other hand, removed any reticence for producing more edgy and opinionated content. This usually happened with no consideration for the bot’s interlocutor, giving no space for real engagement.

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 the model generated, such as in the example above, could get uncomfortable for some. Moreover, the participants pointed out that Leslie-bot tended to go into “rambling mode” without considering the participants’ questions, prompts or responses, thus not allowing for meaningful interactions.

As Leslie-bot clearly signaled a strongly defined personality, participants couldn’t avoid making sense of their interaction with the bot through personification. Leslie-bot stated strong opinions that could be read as forms of disclosure, of making the self vulnerable to the other. Yet as Obert [66] points out, where intimacy is concerned, expressions of vulnerability without the attended curiosity about the other and their opinions can become a form of selfishness, eventually dismissed. Conversations need reciprocity and consideration to create a positive experience no matter the topic [82]. This may be even more important when the topics under discussion are normatively challenging. Although the bot mimics personality (the one of Leslie), neither reciprocity nor consideration are possible for a model, no matter what sort of data it is fine-tuned on. This kind of forced insertion of value-laden notions into conversation, then, is more likely to elicit frustration and eventual dismissal rather than any kind of reflective consideration that could bring forth normative change. After all, nobody likes having opinions forced on them without considering their own point of view and reciprocating expectations of the other part in a conversation, no matter how ostensibly progressively positive.

5 Discussion

5.1 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 efforts such as the European AI Act. Yet as Moats and Ruckenstein [64] 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, one 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 [50]. 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 Leslie-bot 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. Although both you-bot and Leslie-bot were able to mimic human values through fine-tuning, the output they produced was without consideration for the participants they interacted with, unable to reciprocate acknowledgment of their values and belief systems.

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 as well as the context in which the interaction takes place. Returning to Haraway’s argument [39], the LLM’s mimicry of values through fine-tuning takes them out of nowhere, as they still are models without consideration and therefore lacking empathy and reciprocity that would enable relating to the participants. The raging debates around progressive and conservative notions of social values such as abortion, gender, or what constitutes racial justice (to name a few) are examples that demonstrate the complexity of such values. It is no surprise, then, that attempts to “build in” particular values into LLMs can have problematic consequences [76]. Recently, Google (with Gemini) and OpenAI (with GPT-4 and Dall-E 3) have attempted to force these systems to “align” with values of representation and diversity. Gemini and Dall-E 3 edit 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 [48]. Sure, we can say that people simply do not like change, wanting to continue living in our cozy, even if ostensibly ethically problematic worlds and the only way to create change is to force it.

Our work suggests that forcing values in this fashion is potentially counter-productive and that there must be a consideration for how best to approach 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 signaled personality, which essentially forced our participants to instinctively personify the bot. This artificial construction of intimacy developed 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.

The mimicry of vulnerability and reciprocal desire to know and care for the other can lead towards increased engagement with the chatbot [35, 52]. Yet, this desire to know and be known is ultimately impossible for a machine to fulfill, for to know and be known by another, there must be another there; whereas with LLMs, “there is no there there” [8]. Therefore, understanding, care, and knowledge sharing are impossible beyond a one-way human vulnerability to desire these and a returned acknowledgment of that desire [70, 88]. This acknowledgment can be interpreted as an artificial intimacy especially when masquerading as authentically “person-like” [15] as in the case of personalized fine-tuned LLMs that establish trust [51], encouraging the promise of an authentic disclosure on part of the conversational agent that cannot possibly be fulfilled.

5.2 Meaningful conversations: information vs intimacy

Chatbots are designed to mimic human communication, but that 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, and relational closeness. If we take Sigman’s notion of consequentiality of communication seriously [82], then what happens in conversation is sense-making that comes to constitute the social realities we inhabit. Our participants noted that quite often the GPT-3 output felt transactional and robotic, a pertinent representation of the notion that gaining information is the sole purpose of conversation worth consideration. 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 model is already fine-tuned through supervised learning to be as generically palatable as possible. Acknowledging that conversation is a reciprocal process, the GPT 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 an information exchange. It is also a form of world-building. It can be boring, mundane, utterly instrumental [1], and, at times, entirely devoid of content [29] just as much as it can be playful, creative, socially meaningful [53], and full or what Charles Sanders Peirce called “the play of musement” [23]. There is the potential

for conversations to establish intimacy which in itself is a world-building process [11] in that it “creates spaces and usurps places meant for other kinds of relation.” [11]. Intimacy does this within the contexts of shared vulnerability, curiosity, and empathy [66] but with ambivalence towards the achievement of these desires [12]. As Berlant points out, “Contradictory desires mark the intimacy of daily life: people want to be both overwhelmed and omnipotent, caring and aggressive, known and incognito” [11].

It is no wonder that our participants enjoyed the models fine-tuned on personal social media data, whether their own or Leslie Fosters’. The fine-tuned models presented something more emotionally coherent, even if sometimes repetitive, annoying, and bewildering. They could mimic the joy of playful and creative conversation better. In his reflections on the implications of ELIZA, Weizenbaum [91] 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, although it is a proposal that repeatedly comes up in current research on how to make chatbots appear more trustworthy and palatable [51, 52, 71]. Privacy and regulatory concerns are part of this, but this also remains an ongoing technical challenge. People using chatbots often 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 [67]. Implementations such as memory for ChatGPT create opportunities for the chatbot to use familiar terms and turns of phrase, creating potential for 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, providing for the sense-making and world-building that constitutes communication for people, they can only mimic these qualities, opening up possibilities for emotional manipulation by holding up a mirror with nothing of substance behind it.

5.3 Artificial intimacy between capitalism & care

Our research also points to the problematic encounters with pared down, uncontroversial versions of chatbots. People may want a transactional chatbot, service bot, something that completes tasks for them [6], yet this cannot be achieved without erasure by flagging, refusal to respond, vague responses, or by introducing bias and values in seemingly normatively-neutral descriptions and stories. In return, such attempts at normative neutrality 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 or consider, and in fact oppresses their identities and communities [80]. Lingel and Crawford [56] 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 convictions become maintained and developed through the imaginary of the normatively neutral AI assistant, but the underlying processes of inequality, gender politics, data collection systems, and labor need to be protested.

Savolainen and Ruckenstein [79] showcase how intimacy is mimicked by the personalization 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 personalized and emotional connection when relating to these systems. Our participants too were trying to relate to the chatbot through the artificial intimacy created by shared humor, mirroring the self or mimicry of personality. Yet, rather than fostering autonomous agency, these systems push responsibility to the individual, as the systems cannot produce reciprocity and consideration in a given conversation. We need to go beyond the idea of autonomy being created by social and technical others respecting and recognizing us. This cannot be done with artificial intimacy or injecting the ‘right’ values through value alignment. We need systems that go beyond mere technical models, as they have to provide a space, where we can become ourselves in all our diversity [25] and empirically pay attention to the moments when alignment and friction occur in algorithmic systems [79]. This cannot be achieved by erasure of diversity through silencing to the lowest common denominator in an effort to make things non-offensive, as interactions with algorithmic systems need to take place across diversity to increase the actual autonomy of people. Fine-tuning for personalization and normativity makes visible some of these frictions but without overcoming them. With the lack of reciprocity and consideration it can at best mimic care resulting in artificial intimacy. Algorithmic systems incur additional considerations in regards to intimacy as well as potential abusiveness because they lack transparency in the kinds of programming underlying their functionality and they cannot take responsibility for themselves. Instead, they are products that are designed, developed, and deployed by individuals and companies who can opportunistically exploit a mimicry of human intimacy.

6 Conclusion

Chatbots are not a new technology, but advances in LLMs have resulted in better, more articulate, and more engaging chatbots that are 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 of 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 the implications are for 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 the 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 the incorporation of a particular ethics and worldview into the chatbot, gave no room for discussion and reflexivity because the bot could not, by definition, extend 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 arbitrarily selected values can result in normative challenge without consideration, and such challenge is unlikely to be productive.

Acknowledgments

We extend our thanks to Obaida Hanteer for support with the technical aspects of Artificial Intimacy, to Leslie Foster and Gorjeoux Moon for their trust and for allowing us to explore alternative normativities with them. We also thank metaLAB Harvard for hosting the Artificial Intimacy art installation, and all of our workshop participants for persevering despite the technical problems and challenges.

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*. ACM, 1–9.
- [3] Mina Almasi and Anton Schionning. 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] 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*. ACM, 1–8.
- [5] Amanda Askill, 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>
- [6] Joan Palmiter Bajorek. 2019. Voice recognition still has significant race and gender biases. *Harvard Business Review* 10 (2019), 1–4.
- [7] 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.
- [8] 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* (Online). Association for Computational Linguistics, Stroudsburg, PA, USA.
- [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] Lauren Berlant. 1998. Intimacy: A Special Issue. *Critical Inquiry* 24, 2 (1998), 281–288. <https://doi.org/10.1086/448875>
- [12] Lauren Berlant. 2011. *Cruel Optimism*. Duke University Press, Durham, NC. 58 pages.
- [13] Lauren Berlant. 2022. *On the inconvenience of other people*. Duke University Press, Durham, NC.
- [14] Pietro Bernardelle and Gianluca Demartini. 2024. Optimizing LLMs with direct preferences: A data efficiency perspective. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region* (Tokyo Japan). ACM, New York, NY, USA, 236–240.
- [15] Sutcliffe Bianca. 2024. Artificial Intimacy : Exploring Intimacy in Human and AI-enabled Chatbot Relations : its Existence, its Authenticity and its Moral Implications. <http://essay.utwente.nl/101169/>
- [16] Henk Borgdorff. 2010. The production of knowledge in artistic research. In *The Routledge companion to research in the arts*. Routledge, 44–63.
- [17] Consumer Financial Protection Bureau. 2023. Chatbots in Consumer Finance. <https://www.consumerfinance.gov/data-research/research-reports/chatbots-in-consumer-finance/chatbots-in-consumer-finance/>
- [18] Jordi Calvet-Badement and Jacob Mchangama. 2024. AI chatbots refuse to produce 'controversial' output - why that's a freedom of speech problem. <https://theconversation.com/ai-chatbots-refuse-to-produce-controversial-output-why-thats-a-free-speech-problem-226596>.
- [19] Alexander Campolo and Kate Crawford. 2020. Enchanted determinism: Power without responsibility in artificial intelligence. *Engag. Sci. Technol. Soc.* 6 (Jan. 2020), 1–19.
- [20] Caitlin Chin-Rothman and Mishaela Robison. 2020. How AI bots and voice assistants reinforce gender bias. <https://www.brookings.edu/articles/how-ai-bots-and-voice-assistants-reinforce-gender-bias/>.
- [21] 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.
- [22] Kenneth Mark Colby, Sylvia Weber, and Franklin Dennis Hilf. 1971. Artificial paranoia. *Artificial intelligence* 2, 1 (1971), 1–25.
- [23] 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).
- [24] Juliet Corbin and Anselm Strauss. 2014. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications.
- [25] Kate Crawford. 2016. Can an algorithm be agonistic? Ten scenes from life in calculated publics. *Science, Technology, & Human Values* 41, 1 (2016), 77–92.
- [26] Andrea Cuadra, Maria Wang, Lynn Andrea Stein, Malte F Jung, Nicola Dell, Deborah Estrin, and James A Landay. 2024. The illusion of empathy? Notes on displays of emotion in human-computer interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA), Vol. 32. ACM, New York, NY, USA, 1–18.
- [27] Hans Demeyer. 2023. Lauren Berlant on Intimacy as World-Making-interview. *Extra, Extra* 16 (2023), x. <https://doi.org/talk/lauren-berlant-on-intimacy-as-world-making/> Accessed: 2024-12-06.
- [28] Rebecca Dorn, Lee Kezar, Fred Morstatter, and Kristina Lerman. 2024. Harmful speech detection by language models exhibits gender-queer dialect bias. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (San Luis Potosi Mexico). ACM, New York, NY, USA, 1–12.
- [29] 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.

- [30] Catherine D'Ignazio and Lauren F. Klein. 2023. *Data feminism*. The MIT Press.
- [31] Benj Edwards. 2024. Open AI CEO: We may have AI superintelligence in "a few thousand days". <https://arstechnica.com/information-technology/2024/09/ai-superintelligence-looms-in-sam-altmans-new-essay-on-the-intelligence-age/>
- [32] Chris Elsdén, 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.
- [33] Laura Forlano and Itziar Barrio. 2024. From data doubles to data demons: Reflections on a CripTech collaboration. *Leonardo* 57, 2 (April 2024), 132–140.
- [34] Jason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines* 30, 3 (2020), 411–437.
- [35] Nitesh Goyal, Minsuk Chang, and Michael Terry. 2024. Designing for Human-Agent Alignment: Understanding what humans want from their agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu HI USA). ACM, New York, NY, USA.
- [36] 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>
- [37] Güven Güzelçere and Stefano Franchi. 1995. Dialogues with colorful "personalities" of early AI. *Stanford Humanities Review* 4, 2 (1995), 161–169.
- [38] Mike Hannula, Juha Suoranta, and Tere Vaden. 2014. *Artistic Research Methodology: Narrative, Power and the Public*. Peter Lang Publishing.
- [39] 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.
- [40] Will Douglas Heaven. 2022. The new version of GPT-3 is much better behaved (and should be less toxic). <https://www.technologyreview.com/2022/01/27/1044398/new-gpt3-openai-chatbot-language-model-ai-toxic-misinformation/>
- [41] 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
- [42] 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.
- [43] Kristina Hook. 2018. *Designing with the body: Somaesthetic interaction design*. MIT Press.
- [44] Falk Hübnér. 2024. *Method, Methodology and Research Design in Artistic Research: Between Solid Routes and Emergent Pathways*. Taylor & Francis.
- [45] 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/>
- [46] Belén Jiménez-Alonso and Ignacio Brescó de Luna. 2024. AI and grief: a prospective study on the ethical and psychological implications of deathbots. In *Ethics in Online AI-based Systems*. Elsevier, 175–191.
- [47] 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>
- [48] Jereny Kahn. 2024. *What Google's 'woke' AI image controversy says about AI—and about Google*. <https://fortune.com/2024/02/27/google-gemini-woke-ai-images-alphabet-sundar-pichai/>
- [49] 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.
- [50] 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>
- [51] Nikola Kovacevic, Tobias Boschung, Christian Holz, Markus Gross, and Rafael Wampfler. 2024. Chatbots with attitude: Enhancing chatbot interactions through dynamic personality infusion. In *ACM Conversational User Interfaces 2024* (Luxembourg Luxembourg). ACM, New York, NY, USA.
- [52] Nikola Kovacević, Christian Holz, Markus Gross, and Rafael Wampfler. 2024. The personality dimensions GPT-3 expresses during human-chatbot interactions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 2 (May 2024), 1–36.
- [53] Cassandra Lee and Jessica R Mindel. 2024. Closer and closer worlds: Using LLMs to surface personal stories in world-building conversation games. In *Designing Interactive Systems Conference* (IT University of Copenhagen Denmark). ACM, New York, NY, USA.
- [54] Davide Liga and Livio Robaldo. 2023. Fine-tuning GPT-3 for legal rule classification. *Computer Law & Security Review* 51 (2023), 105864.
- [55] Nora Freya Lindemann. 2022. The ethics of 'deathbots'. *Sci. Eng. Ethics* 28, 6 (Nov. 2022), 60.
- [56] 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).
- [57] 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>
- [58] Maia Mackney and Toby Young. 2022. Embracing the messiness: a creative approach to participatory arts evaluation. *Cultural Trends* 31, 5 (2022), 397–415.
- [59] Takuya Maeda and Anabel Quan-Haase. 2024. When human-AI interactions become parasocial: Agency and anthropomorphism in affective design. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro Brazil). ACM, New York, NY, USA.
- [60] 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.
- [61] Dave McKean. 2022. *Black Dog: The Dreams of Paul Nash*. Dark Horse Comics.
- [62] Microsoft. 2024. *Data, privacy, and security for Azure OpenAI Service*. <https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy>
- [63] Niloofar Miresheghallah, Maria Antoniak, Yash More, Yejin Choi, and Golnoosh Farnadi. 2024. Trust No Bot: Discovering Personal Disclosures in Human-LLM Conversations in the Wild. In *First Conference on Language Modeling*. <https://openreview.net/forum?id=tlpWtMYkzU>
- [64] 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*.
- [65] Mohammad Namvourpour and Afsaneh Razi. 2024. Uncovering contradictions in human-AI interactions: Lessons learned from user reviews of replika. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing* (San Jose Costa Rica). ACM, New York, NY, USA, 579–586.
- [66] Julia C. Obert. 2016. What we talk about when we talk about intimacy. *Emotion, Space and Society* 21 (2016), 25–32. <https://doi.org/10.1016/j.emospa.2016.10.002>
- [67] OpenAI. 2024. *Memory and new controls for ChatGPT*. <https://openai.com/index/memory-and-new-controls-for-chatgpt/>
- [68] 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) (AI/ES '23). Association for Computing Machinery, New York, NY, USA, 496–511. <https://doi.org/10.1145/3600211.3604705>
- [69] Ozlem Ozmen Garibay, Brent Winslow, Salvatore Andolina, Margherita Antona, Anja Bodenschatz, Constantinos Coursaris, Gregory Falco, Stephen M Fiore, Ivan Garibay, Keri Grieman, et al. 2023. Six human-centered artificial intelligence grand challenges. *International Journal of Human-Computer Interaction* 39, 3 (2023), 391–437.
- [70] Susanna Paasonen. 2017. Infrastructures of intimacy. In *Mediated intimacies*. Routledge, 103–116.
- [71] Pat Pataranutaporn, Ruby Liu, Ed Finn, and Pattie Maes. 2023. Influencing human-AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence* 5, 10 (2023), 1076–1086.
- [72] John Durham Peters. 1999. *Speaking into the Air: A History of the Idea of Communication*. University of Chicago Press.
- [73] William Chamberlain Racter. 1984. *The Policeman's Beard is Half Constructed: Computer Prose and Poetry by Racter - The First Book Ever Written by a Computer*. Grand Central Pub.
- [74] 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.
- [75] Juan Rebanal, Jordan Combitis, 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.
- [76] 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>
- [77] Emma Roth. 2024. *Character.AI and Google sued after chatbot-obsessed teen's death — theverge.com*. <https://www.theverge.com/2024/10/23/24277962/character-ai-google-wrongful-death-lawsuit>
- [78] Minna Ruckenstein. 2023. *The feel of algorithms*. Univ of California Press.
- [79] Laura Savolainen and Minna Ruckenstein. 2024. Dimensions of autonomy in human-algorithm relations. *New Media & Society* 26, 6 (2024), 3472–3490.
- [80] 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.

- [81] Britta Schulte and Eva Hornecker. 2020. Full Frontal Intimacy - on HCI, Design & Intimacy. In *Companion Publication of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven Netherlands). ACM, New York, NY, USA.
- [82] Stuart J Sigman. 2013. Introduction: Toward study of the consequentiality (not consequences) of communication. In *The consequentiality of communication*. Routledge, 1–14.
- [83] 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.
- [84] 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.
- [85] 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.
- [86] The Brussels Times. 2023. Belgian man dies by suicide following exchanges with chatbot – brusselstimes.com. <https://www.brusselstimes.com/430098/belgian-man-commits-suicide-following-exchanges-with-chatgpt>.
- [87] Sherry Turkle. 2020. A nascent robotics culture: New complications for companionship. In *Machine ethics and robot ethics*. Routledge, 107–116.
- [88] Sherry Turkle. 2024. Who Do We Become When We Talk to Machines? *An MIT Exploration of Generative AI* (mar 27 2024). <https://mit-genai.pubpub.org/pub/uawlh3j>.
- [89] 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>
- [90] Joseph Weizenbaum. 1977. Computers as "Therapists". *Science* 198, 4315 (1977), 354–354.
- [91] Joseph Weizenbaum. 1995. The myth of the last metaphor. *Speaking minds: Interviews with twenty cognitive scientists* (1995), 249–264.
- [92] 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.
- [93] Langdon Winner. 2009. Do artifacts have politics. *Readings in the Philosophy of Technology* 251 (2009).
- [94] 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.
- [95] Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying Language Models Risks Marginalizing Minority Voices. <https://aclanthology.org/2021.naacl-main.190/>
- [96] Aijia Yuan, Edlin Garcia Colato, Bernice Pescosolido, Hyunju Song, and Sagar Samtani. 2024. Improving workplace well-being in modern organizations: A review of large Language Model-based mental health chatbots. *ACM Trans. Manag. Inf. Syst.* (Oct. 2024).