Approaches to Epistemic Risk in Generative and General-Purpose AI

Robert Wolfe

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:
Bill Howe, Chair
Alexis Hiniker, Chair
Tanu Mitra

Program Authorized to Offer Degree:
Information School

University of Washington

Abstract

Approaches to Epistemic Risk
in Generative and General-Purpose AI

Robert Wolfe

Co-Chairs of the Supervisory Committee:
Bill Howe
Alexis Hiniker
Information School

Generative and general-purpose AI systems stand poised to reshape longstanding information infrastructures and professions, ranging from search to social media to online journalism. Yet questions surrounding subtle biases, misinforming output, and system reliability and transparency – epistemic risks related to the way knowledge is encoded and disseminated – have followed these technologies since their inception. Without strategies for understanding and managing the risks they pose, general-purpose models may degrade the reliability of the information ecosystem, as well as introduce hazards for the individuals and institutions deploying them. This dissertation introduces methods to understand epistemic risks in generative and general-purpose AI and approaches to responsibly deploy these systems in the presence of inevitable epistemic risk.

Concretely, this dissertation develops three approaches to epistemic risk in generative and general-purpose AI. First, I introduce computational approaches to identifying both the manifestations of epistemic risks like bias and misrepresentation and their underlying causes, such as the scale of a model's pretraining dataset and the unanticipated biases present in high-quality media data such as online newspaper articles. Second, I introduce novel design frameworks that account for epistemic risk in generative models, taking into account the need for information integrity among organizations engaged in data-driven knowledge work, as

well as among users in interpersonal communication online. Finally, I introduce transparency-maximizing approaches to mitigate the heightened epistemic risk of using generative models served over black-box APIs, including an approach that customizes small open models on consumer-grade GPUs, as well as a context-sensitive approach to the adoption of open and proprietary models that accounts for the needs of organizations engaged in human-centered data science work. Taken together, these approaches point toward a future for generative and general-purpose AI that values reliability and information integrity.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am glad to have the opportunity to express gratitude to the many people and institutions who have shaped and supported me on the path to the Ph.D.

This must begin with my *co-advisors*, **Dr. Alexis Hiniker** and **Dr. Bill Howe**. Despite my relative lack of experience in HCI, Alexis took me into her User Empowerment lab and provided the kind of constant mentorship that has allowed me to grow into a mature scientist. Her qualitative research and career preparation classes provided some of the most important resources for setting me on the path I now pursue. Bill trusted me with leading the early work of his Volitional AI lab, producing one of the studies included in this dissertation, and his guidance has helped me to center responsible and transparent science in my own research. His confidence in me has often exceeded my own, and reflects his unusual capacity to see the best in people. Both Bill and Alexis have equipped me with perspective that will stay with me long beyond the Ph.D.

Beyond my advisors, I have benefited from the guidance of many *mentors* throughout the Ph.D. **Dr. Tanushree Mitra** advised two of the most interesting studies I've pursued so far, and she provided direct, concrete advice that improved the rigor of my qualitative research. **Dr. Leilani Battle** has generously served as the GSR of my supervisory committee, and she has provided consistently insightful and helpful feedback at all of the major milestones of my Ph.D. journey. **Dr. Nic Weber** has expanded the intellectual richness of my Ph.D. experience with his insights about novel research methods and adjacent subfields of information science. My first teaching experience came with **Dr. Jevin West**, who gave me the opportunity to develop course materials and teach several class sessions in the second year of my Ph.D., a critical early experience that built my confidence as a new instructor. **Dr. Wanda Pratt**

served on my advisory committee during the early years of my Ph.D., and she generously welcomed me to her lab during a challenging time of transition. **Dr. Noah Smith** and **Dr. Yulia Tsvetkov** also served on my advisory committee during the first year of my Ph.D., and I am grateful for their time and guidance. I am also grateful to **Dr. Clarita Lefthand-Begay** and **Nicole Kuhn** for welcoming me into their directed research group during the second year of the Ph.D. Several of the most enjoyable experiences of the Ph.D. have come at the Kids Team program with **Dr. Jason Yip**, where I've had a chance to see kids interact, test, and often joyfully break the technologies I've designed with other researchers. I'm thankful to **Dr. Anind Dey** for his capable leadership of the iSchool. Finally, I remain grateful to **Dr. Robert Pless** and **Dr. Abdou Youssef** of the George Washington University, who served on my thesis committee on very short notice during the summer of 2021, allowing me to rapidly complete my master's degree at GWU and begin the Ph.D. at UW.

The support of the iSchool's *institutional staff* has also been important for my success during the past few years. **Nick Dempsey** was a godsend, providing technical guidance at numerous times during the Ph.D., and extraordinary assistance after a hard drive failure that left me scrambling to retrieve seemingly lost data. **Leigh Eisele** and **Andrea Gnessin** have made the iSchool a warmer place, and offered a receptive ear to me and to many other students. I also am grateful to **Dowell Eugenio** for helping me better understand the iSchool during my early days on campus. **Megan Greene**, **Jon Larson**, and **Nicky West** have provided invaluable support for obtaining the technical infrastructure underlying many of the studies in this dissertation. The consistent competence of **Lily Allred**, **Kate Kerschbaum**, **Matt Saxton**, and **Wendie Phillips** has made my path through the Ph.D. relatively seamless.

I have had the good fortune to be surrounded by many *peers* who embody the excellence of the iSchool. This is certainly true of the **i'21 cohort**, whose diverse approaches to information

science have helped to expand my understanding of how to do great research. I am particularly grateful to the members of Dr. Howe's **Volitional AI Lab**, including the many collaborators on the laboratory-scale AI project, and to the members of Dr. Hiniker's **User Empowerment Lab**. These groups have provided me with a vibrant research community during the Ph.D. I am especially grateful to those peers who have invited me to contribute to their research. These include **JaeWon Kim**, **Ishita Chordia**, **Amanda Baughan**, **Lucas Rosenblatt**, **Bingbing Wen**, **Yiwei Yang**, **Bin Han**, and **Zening Qu** — and, of course, **Aayushi Dangol**. I have learned so much from all of you.

Though I will graduate as an information scientist, I began my academic life as a student of literature, and I am deeply grateful to the many *humanists* whose mentorship guided my early life and ultimately paved the path to scientific inquiry. **Dr. Ingrid Satelmajer** taught me to see the points of intersection between science and art, a perspective that has proven beneficial throughout my graduate studies. **Dr. Matthew Pavesich** helped me arrive at the intellectual foundation for my first master's thesis, and he inspired an interest in rhetorical ecologies that ultimately led to my study of NLP and information science. **Dr. Brian Hochman** introduced me to media studies, a field that has remained central for me throughout the Ph.D. **Professor John Auchard** taught me the importance of paying close enough attention to notice the meaning even of those things that go unsaid. **Dr. Norma Tilden** reminded me that academics should not *guard* the door to the ivory tower, but find ways to open it ever wider. **Dr. Ricardo Ortiz** helped me to understand the real-world impacts that can be achieved through humanistic scholarship. **Maud Casey**, **Tom Earles**, **Noah Siela**, **Johnna Schmidt**, and **Dean Hebert** fostered my abilities as a writer and an artist, and helped me understand that sometimes one can only approach the truth by writing one's way into it.

It is also safe to say that I would never have had the opportunity to pursue the Ph.D. without the guidance of my *colleagues* at DisputeSoft, where I spent nearly seven years of my

professional life. **Jeff Parmet** provided the best example of capable, responsible leadership anyone could hope for, and I'm deeply grateful for the years I spent learning from him. **Todd Trivett** became a great friend, mentor, and civic-minded fellow traveler during the COVID-19 pandemic. **Brendan McParland** reshaped the way I write, and his presence can likely still be felt in many of the stylistic choices made throughout this dissertation. **Nick Ferrara** could always be counted on for a lively conversation, and he did much to further my early computer science education by sketching out proofs on our office whiteboards as I took classes in the evenings. **Josh HelfinSiegel** exemplified the consistency and good character that I aspire to as a professional, and he taught me my way around computer hardware. I was never able to stop saying a midwestern goodbye to **T.J. Wolf**, who I'm glad to have been able to see a few times during the Ph.D. **Hunter Jones**, **Frank Hydoski**, and **Allen Klein** shared with me the kind of wisdom that can only come from a long, successful career. **Mikaela Berst**, **Haley Miller**, **Amanda Doran**, **Anne Ackerman**, **Tom Ashley**, **Sandelle Sefa**, **Aparna Kaliappan**, and **Evan D'Aversa** were great colleagues throughout my years at the firm.

I am also grateful to the *friends* who have stuck with me through the Ph.D. These include **Jean Salac** and **John Zhang**, who changed the course of my life by introducing me to Aayushi at a potluck, and who have remained constant and devoted friends in the years since. I am glad to have met **Kunsang Choden** during the Ph.D., whose good energy has often lifted my spirits. I am also happy to have found friendship with **Charles Bugre**, **Turam Purdy**, **Chris Fu**, **Isaac Slaughter**, **Katelyn Mei**, and many other fellow Ph.D. students who made UW a happier place throughout the Ph.D. **Tarsilla Moura**, **Taylor Osbourne**, and **Sohayl Vafai** have been some of my oldest friends, and I'm grateful for the times we've been able to see each other during the Ph.D., despite the distance. I'm also grateful to the old friends who have made an effort to reach out from afar, including **Tori Peck**, **Tori Kronz**, and **Rebecca Ogle**. Finally, I'm grateful to the **Fairfax Freethinkers** for the rich

intellectual community I found a home in for several years, including remotely during the pandemic and through the beginning of the Ph.D.

I am deeply grateful to my *family*. I am the beneficiary of decades of hard work on the part of my parents, **Robert Wolfe, Jr.** and **Sue Wolfe**, and I am grateful for their love and support during the Ph.D. I am also grateful for my brother, **Steven Wolfe**, and his wife **Chrissa Wolfe**, and I look forward to being close enough to drive to West Virginia again soon. It has also been wonderful to become part of a new family during the Ph.D. Exploring Seattle with my father-in-law **Purushottam Dangol** was one of the most enjoyable weeks of the past few years. I'm also glad to have experienced a little bit of Nepal over FaceTime with my mother-in-law **Ishwari Dangol** and sister-in-law **Abhigya Dangol**.

Finally, I am grateful to my *wife*, **Aayushi Dangol**. She has been the greatest companion, friend, and collaborator I could have asked for since the day we met, and I would not have finished this dissertation without her. Though I will miss the wonderful home we have made together in Seattle, I also look forward taking the next steps toward building a life together on the east coast.

DEDICATION

For Aayushi.

Thesis Statements

This dissertation makes six claims concerning epistemic risk in generative and general-purpose AI.

- **Statement 1**: Epistemic risk in generative and general-purpose AI models results not only from explicitly toxic training data but from using ethical, high-quality data sources outside of their original context in order to train AI.

- **Statement 2**: Epistemic risk in general-purpose models is predicted by 1) the size of the dataset on which a model has been pretrained and 2) the relative consistency of the bias, insofar as this can be captured by large-scale psychometric surveys.

- **Statement 3**: Design that centers the verification of information can mitigate epistemic risk when using generative and general-purpose AI in the production of knowledge.

- **Statement 4**: Design that centers attunement to human needs can mitigate epistemic risk when using generative and general-purpose AI in interpersonal communication.

- **Statement 5**: The heightened epistemic risks posed by closed-weight proprietary models can be mitigated by fine-tuning small open models to perform specific tasks while maintaining their general-purpose chat interface.

- **Statement 6**: In practice, human-centered data science work benefits from a blend of open and proprietary models that leverages the strengths of each paradigm to reduce epistemic risk.

Chapter 1

INTRODUCTION

General-purpose AI systems like the now-ubiquitous ChatGPT stand poised to reshape longstanding information infrastructures and professions, ranging from search to social media to online journalism [127, 242]. Yet questions surrounding subtle biases, misinforming output, and uncertain data provenance—*epistemic risks* related to the way knowledge is encoded and disseminated—have followed these technologies since their inception [28, 424]. Understanding the nature of these risks and managing them effectively in models that can take conceivably any input and produce a plausible output remains a complex and high-stakes challenge that necessitates approaches not only computational but also social and qualitative.

The primary contributions of this dissertation describe approaches to 1) identify and characterize the manifestations of epistemic risks like bias and misinformation in generative and general-purpose AI, along with their underlying causes; 2) develop context-sensitive techniques and designs that preserve information integrity, allowing for the responsible deployment of epistemically flawed but nonetheless useful technologies; and 3) produce alternative approaches to closed, proprietary systems, mitigating the heightened epistemic risk of these models where reliability and transparency are valued as highly as raw performance.

1.1 Characterizing Emergent Epistemic Risk in Generative and General Purpose AI

Impressive technological advances in general-purpose AI are often marred by incomplete or misguided assessments of epistemic risk. What's more, where such systems fail, they often do so by undermining the very values their organization intended to promote. Meta's Galactica, a model intended for generating scientific articles [386], was shut down 72 hours after launch

due to its proclivity to generate misinformation, authoritatively presented in the syntax of scientific communication [168]. Google's Gemini image generator, a model trained to boost the representation of minority racial and gender groups [387, 336], produced images of people of color when prompted to generate images of Nazis, depictions that were not only offensive but also disturbing to users [341]. When it comes to managing epistemic risks in systems built on large-scale web scrapes, good intentions—and partial evaluation—are not sufficient.

Managing the risk of bugs or downtime in traditional software systems meant writing tests to ensure that deployed software fulfilled its intended purpose and met standards of reliability expected by consumers [401]. Yet the capacity of general-purpose models to take nearly any input and produce a plausible output presents a problem that is as much social as technical, as AI may replicate subtle, even unconscious biases of the human mind [68, 67]. These problems can affect even production-grade models [19], impacting systems intended to perform societally beneficial functions. Consider BeMyAI, a system using OpenAI's GPT-4-**V**ision model to describe the world in real-time to low-vision users [24]. Though broadly useful, OpenAI disabled the system's capacity to describe *people* because it made unwarranted inferences about human emotions and character traits (such as trustworthiness) based on facial features [173]—a problem known "facial impression bias" [396].

Though studied by psychologists using methods like three dimensional mesh modeling [48], facial impression bias had never before been documented in general-purpose AI models, which are not explicitly trained to make such inferences. Chapter 4 of this dissertation, however, provides evidence for precisely this problem, as I used a dataset developed by psychologists [304] to study a suite of 43 pretrained multimodal CLIP models (classifiers that learn to match images with text based on the cosine similarity between them [321]), with systematically varying training regimes and parameter counts [84]. I found that two variables predominantly accounted for the traits learned by the models: the extent to which the inference was consistent across society (based on human inter-rater reliability), and the size of the dataset on which the model trained. This introduces a catch-22 for training general-purpose AI: the same variables that produce models that are more accurate (scale) and

preferred by more people (societal consistency) also lead to novel biases, with consequences for downstream applications.

Moreover, aiming for models that appeal to the broadest base of people can reproduce skewed representations of social groups reflective primarily of their depiction in the popular media. Teenagers, for example, are a vulnerable group in many societies who lack agency due to their age, and a large body of research establishes that popular and news media tends to present teenagers as either *a risk* to society or *at risk* from society, with responsible or civic-minded teenagers presented as notable exceptions to the norm [292, 373, 17]. Chapter 3 of this dissertation shows that these stereotypes exist not only in these media sources but also in the AI models trained on them, as teenagers are represented as rebellious, violent, and sexually vulnerable, both in static word embeddings and in modern generative models [440]. Because prior work describes differences in the representation of teenagers across cultures [212, 112], I also studied whether AI trained on Nepali text corpora reflect similar biases, working with a primary co-author who was a native speaker of Nepali. While not free of stereotypes about adolescents, Nepali embeddings and generative models exhibited little of the lurid sensationalism present in English models.

Addressing these problems requires actually understanding the perspectives of the populations thus represented. To do so, I held workshops with N=14 English-speaking adolescent participants residing in the U.S. and N=18 Nepali-speaking adolescent participants residing in Nepal. The participants received a brief tutorial about how AI is trained, and were then asked for their perspectives about how teenagers were represented in the media, and how they should be represented in AI. English-speaking participants foregrounded that AI should represent the *diversity* of teenagers, while Nepali-speaking participants foregrounded that AI should represent teenagers with *positivity*. Managing epistemic risk that affects a specific, vulnerable population requires more than a one-size-fits-all approach, pointing to the need to develop models that can be personalized to the needs of smaller populations, rather than to the needs of the average user.

The studies above continue a line of work in which I have applied and developed psy-

chologically grounded tests for quantifying epistemic risk in general-purpose AI systems [438, 436, 442], especially multimodal models like CLIP and Stable Diffusion [433, 434, 446]. My research on these models has studied consequential topics ranging from biased defaults in general-purpose classification [437] to problems of safety when generating images of women [446], highlighting unanticipated epistemic problems in novel AI architectures.

1.2 Designing for Information Integrity in the Presence of Epistemic Risk

Flaws that prevent general-purpose AI from reaching its full potential do not necessarily prevent its responsible use today; BeMyAI can still provide a remarkable interface to most of the world for low-vision users, even if it is programmatically prevented from describing images of people. Thus, while developing effective technical solutions to epistemic issues remains an important research direction, and one I have explored in my collaborations [454], an equally important direction explores how organizations can responsibly deploy general-purpose AI *in the presence of epistemic risk*.

Addressing this question requires studying not only general-purpose AI but the organizations that adopt it. To that end, I conducted an interview study with fact-checking organizations, for whom success depends on navigating epistemic risk not only in their own organization but in society as a whole. To capture global perspectives, I interviewed 38 participants at 29 signatory organizations of the International Fact-Checking Network (IFCN) [312] in 19 countries, asking how interviewees use generative AI, how they envision using it, and what prevents them from adopting it further.

The interviews surfaced tensions between the need to provide audiences with an efficient response to misinforming content circulating online and the need to carefully verify all published fact-checking content. Fact-checking organizations necessarily embrace task-specific AI and machine learning solutions to sift through enormous quantities of online content [195], and most expressed openness to adopting general-purpose AI not only for processing data internally but also for user-facing applications, like collecting tips and delivering fact-checking content via novel interfaces like conversational tiplines. Yet most organizations were also

acutely aware of the risks of a mistake. Even as his organization moved forward with a variety of chatbot-driven solutions, one of the CEOs I interviewed said that an AI-generated error could be a catastrophic event that jeopardizes the organization's existence and casts doubt on its commitment to its expressed values. In light of such risks, how could this organization move forward with deploying generative AI?

The answer is an approach I call **Designing for Verification** [443]. Interviewees consistently described a process of information *production* followed by *verification* to ensure the epistemic integrity of their organization's content. Depending on the context, either a human fact-checker or a generative model could play the role of the Producer and/or the role of the Verifier. Though human fact-checkers did verify machine-generated content (and systematically tested in-beta conversational tiplines), the more common use of generative models (and more appealing, to many organizations) was in quality assurance—fact-checkers valued an adversarial analysis of their own content prior to publication, which helped them to *address epistemic errors* and uphold their organization's values, rather than creating errors that could harm their organization.

Though generative models carry clear epistemic risks for knowledge work, they also have the potential to impact human interactions and relationships by affecting the integrity of *interpersonal* information. In an advertisement for the 2024 Paris Olympics, Google suggested that its Gemini language model might be appropriately deployed to write a letter to an athlete on behalf of an admiring child [150]. The ad, awarded a "gold medal for worst Olympic ad" by *The Atlantic* columnist Caroline Mimbs Nyce [268], was widely criticized for lauding the hollow connection achievable when using AI to produce the contents of human communication [256, 176].

How, then, should AI—and technology more broadly—be used to mediate human connection and communication, if at all? To answer this question, I enrolled N=13 participants in a diary study that examined their empathy-seeking behaviors online. I also conducted an interview study with N=14 certified trainers of Nonviolent Communication (NVC), an approach to interpersonal communication that centers the clear expression and satisfaction

of human needs [347, 346]. Triangulating between the needs of lay users and the advice of NVC trainers yielded an approach to technology in interpersonal settings that I call *Needs-Conscious Design*, which is characterized by its three design objectives of supporting precise observation of needs, supporting taking personal responsibility for needs, and supporting intentional action to meet needs. One of the design risks identified in the study found that positioning AI *between* people often produced an effect I call *empathy fog*, where an individual is unsure of whether the connection they are being offered comes from another person, or from an AI mediator. On the other hand, many participants also proposed designs to help users to slow down their online communication, often with the help of AI, allowing reflect on their intended communication and achieve greater presence with other people. As with knowledge work, human-centered design can help to mitigate the epistemic risk of AI in interpersonal communication.

1.3 Contending With the Heightened Epistemic Risks of Closed, Proprietary Models

While one contends with epistemic risk when using any present generative or general-purpose model, the risk is heightened when using *proprietary* models in particular, like OpenAI's ChatGPT and Anthropic's Claude. Submitting materials through the web-based APIs available for using these models may result in that data being retained by a corporation and used for a purpose not intended by the user, such as training a future iteration of the corporation's language model. Open-source and open-weight models like Mistral [191] or Meta's LLaMA series [398] seem to offer an alternative by allowing companies to download and run models either locally or on a private cloud instance. But how competitive are open models with proprietary models served via corporate APIs, and what tradeoffs do organizations see in using them? I addressed these questions in a study that found that open models like LLaMA-2 lag behind proprietary models like ChatGPT out of the box, but can match the task-specific performance of proprietary models with only a small amount of supervised fine-tuning [445]. The study observed a tradeoff between cost and speed at

runtime: while customized open models could be run at a significant savings on low-cost hardware, they could not match the speed of highly optimized proprietary APIs like that offered by OpenAI. Subsequent experiments showed that proprietary models also produced less misinforming content by default than open models, but the gap could be closed with modest fine-tuning.

Of course, whether organizations choose to use open models or proprietary models is determined by more than technical characteristics like task performance and runtime cost. To better understand the tradeoffs of using open vs. proprietary models, I conducted a study with 24 fact-checking organizations [444], asking specifically about when they used open vs. proprietary models. The study revealed that organizations can't be neatly divided those that use proprietary models and those that use open models; rather, most interviewees saw the tradeoff between open and proprietary models as situation-specific. Open models were preferred when dealing with valuable or sensitive data, or when performance needed to be especially strong for specific tasks, as open models can be extensively customized. Proprietary models were preferred for user-facing applications, as they offer reliable out-of-the-box usability and fairness mitigations that interviewees viewed as safer for audiences. Interviewees also observed that building user-facing applications with proprietary models offered business opportunities that open models precluded, such as integration of a branded chatbot into a proprietary AI ecosystem like OpenAI's GPT Store [285]. Should such ecosystems become more popular, interviewees believed they may serve as a new means of disseminating content and generating revenue. In many cases, the epistemic risks of general-purpose AI can be better managed with a contextual, human-centered approach, rather than a blanket policy governing the use of open or proprietary models.

1.4 Dissertation Overview

This dissertation presents six finished research projects on epistemic risks in generative and general-purpose AI. The projects employ quantitative, qualitative, and mixed methods, as demanded by the research questions animating them. Five of the projects were published in

2024 at the two flagship conferences of the AI Ethics community: the ACM Conference on Fairness, Accountability, and Transparency (FAccT), and AI, Ethics, and Society (AIES). One study remains under submission.

In the second chapter, I review the related work on the subjects germane to this dissertation: the technical and sociotechnical characteristics of generative and general-purpose AI; the reflection of societal attitudes and humanlike biases in these systems; the epistemic risks posed by such systems; and the existing human-centered approaches and designs that enable both the use of AI systems and the study of their epistemic flaws.

In the third chapter, I demonstrate the epistemic risk of training general-purpose models on media data taken out of its original context: namely, a disconnect between the representations learned by these models and the perceptions and experiences of vulnerable and marginalized people. Chapter 3 thus presents evidence for **Thesis Statement 1**: *Epistemic risk in generative and general-purpose AI models results not only from explicitly toxic training data but from using ethical, high-quality data sources outside of their original context in order to train AI.* This chapter reflects the findings of the research project *Adolescent Representational Bias: A Bilingual, Bicultural Study* [440], for which I held workshops with N=13 adolescents in the U.S. and N=18 adolescents in Nepal to understand how they believed they were represented in the media, and how they wanted to be represented in emerging generative AI systems. The views shared by these participants showed that the (often sensationalistic) depiction of adolescents in generative models is misaligned with adolescent lived experiences. Participants expressed two prevailing beliefs about how adolescents should be represented in generative models: Nepalese participants mostly preferred that AI represent them *positively*, while U.S. participants preferred that AI highlight the *diversity* of adolescents.

In the fourth chapter, I show that training general-purpose vision-language AI on ever-larger datasets produces novel epistemic risks, even as it improves the accuracy of such models on benchmarks and traditional machine learning tasks [84]. Chapter 4 presents evidence for **Thesis Statement 2**: *Epistemic risk in general-purpose models is predicted by 1) the size of the dataset on which a model has been pretrained and 2) the relative consistency of*

the bias, insofar as this can be captured by large-scale psychometric surveys. This chapter reflects the findings of the research project *Dataset Size and Human Agreement Mediate Facial Impression Bias in Language-and-Image AI* [439], in which I study the CLIP models trained by Cherti et al. (2022) [84], who demonstrate a scaling law relating CLIP performance to total compute. My results show that associations of *unobservable* traits (*e.g.*, trustworthiness) with images of human faces more closely reflect real-world human bias (as approximated by participants in a large scale online study rating such faces [304]) as dataset scale increases. Moreover, measures of human inter-rater reliability in human facial judgments predict how strongly CLIP biases reflect human biases, indicating that CLIP learns to rely on the same biased heuristics as humans [396, 304] as it observes greater quantities of human data. Facial impression bias plays a socially detrimental role in domains such as hiring [374, 382], criminal sentencing [431], and policing [196]. That models often presented as possessing superhuman visual abilities [61] might *confirm* the bias of a human user compounds the epistemic risk associated with their use.

In the fifth chapter, I show that human-centered design building on the design space of generative models [255] can enable the reliable use of even epistemically flawed models. Chapter 5 presents evidence for **Thesis Statement 3**: *Design that centers the verification of information can mitigate epistemic risk when using generative and general-purpose AI in the production of knowledge.* This chapter reflects the findings of *The Impact and Opportunities of Generative AI in Fact-Checking* [443], for which I conducted an interview study with N=38 professional fact-checkers about the implemented and envisioned uses of generative AI in fact-checking, as well as when and why fact-checkers avoid the technology. I employ a Technology-Organization-Environment (TOE) framework [313] to specify the model-related, organizational, and societal concerns of participants. The study introduces the dimension of Verification, a novel dimension in the design space of generative AI introduced by Morris et al. (2023) [255]. Verification is conceptualized as a four-quadrant space, with the (human or AI) Producer of information on the X-axis, and the (human or AI) Verifier of information on the Y. Participants envisioned designs belonging to all four quadrants to support the

sociotechnical objectives of fact-checking, but they always required that any published content be reviewed by a human Verifier.

In the sixth chapter, I show that human-centered design can facilitate information integrity not only in factual contexts, but in interpersonal contexts, where users' trust in *each other* can depend on the way in which AI mediates human communication. Chapter 6 presents evidence for **Thesis Statement 4**: *Design that centers attunement to human needs can mitigate epistemic risk when using generative and general-purpose AI in interpersonal communication.* This chapter reflects the findings of *Toward Needs-Conscious Design: A Design Framework for Nonviolent Communication*, for which I conducted an interview study with N=14 certified Nonviolent Communication (NVC) trainers [347, 346, 136] and a diary study and co-design with N=13 lay participants, who recorded emotional needs they satisfied via online interaction. This study introduces Needs-Conscious Design, a framework for envisioning technologies that support precise observations of interpersonal needs, encourage users to take personal responsibility for their own needs, and support users in taking intentional action to meet those needs. Needs-Conscious Design employs three levels of attunement to help designers consider whether a technology helps with *self-attunement*, *others-attunement*, or *context-attunement*, the last of which reflects a level of maturity that allows one to prioritize between needs depending on context. Though this study touches on a broader suite of technologies, generative and general-purpose AI were at the front of the mind for most of the participants, and many of the design concepts and design risks produced by the study focus specifically on implementations of generative AI. The study is optimistic about the possibility of AI helping to improve the clarity of human interpersonal communication, but it also clearly identifies the risks of even well-intended applications of generative models in interpersonal communication. Among these risks is empathy fog, a scenario wherein an individual becomes unsure of whether the empathy they are receiving from a conversation partner is genuine, or if it has been manufactured in a low-effort manner using a generative AI writing tool. This study makes clear that, whatever the opportunities it presents, in some cases inserting generative AI to produce more palatable and even more accurate information nonetheless

degrades the value of that information, as well as the human experience underlying it.

In the seventh chapter, I propose an approach to mitigate the heightened epistemic risks posed by closed, proprietary AI, as described in the many recent position papers [293, 243], literature reviews [28, 44], and introductions for technologies for increasing the usability of open models [106, 418, 447]. Chapter 7 thus presents evidence for **Thesis Statement 5**: *The heightened epistemic risks posed by closed-weight proprietary models can be mitigated by fine-tuning small open models to perform specific tasks while maintaining their general-purpose chat interface.* This chapter reflects the findings of *Laboratory-Scale AI: Open-Weight Models are Competitive Even in Low-Resource Settings* [445], in which I investigate whether scientists and public interest organizations can extricate themselves from dependence on proprietary models by leveraging available hardware. The findings are promising: quantizing [108] small-scale open chat models like Mistral-7B-Instruct [191] and LLaMA-2-Chat-7B [399] and using low-rank adaptation [181, 107] to fine-tune them on consumer-grade GPUs produces domain-specialist models that surpass the performance of GPT-4-Turbo, without degrading the chat interface that renders generative models usable by non-experts. The total cost of both fine-tuning and evaluation is less than evaluation alone in GPT-4-Turbo, and only a few hundred training examples are required. Results also empirically validate benefits such as differentially private gradient noising [457, 456] and tunable abstention properties [365, 483] in open models. Choosing open models for reproducible science and data privacy is possible and increasingly practical.

In the eighth chapter, I show that open models present their own epistemic risks and demonstrate that context-sensitive selection of proprietary and open models provides organizations with an effective approach to managing epistemic risk. Chapter 8 presents evidence for **Thesis Statement 6**: *In practice, human-centered data science work benefits from a blend of open and proprietary models that leverages the strengths of each paradigm to reduce epistemic risk.* This chapter reflects the findings of *The Implications of Open Generative Models in Human-Centered Data Science Work: A Case Study with Fact-Checking Organizations* [444], for which I conducted an interview study with N=24 data scientists at fact-checking

organizations. In addition to concerns about the performance, usability, and safety of open models, participants worried about the opportunity cost of choosing open models, anticipating an economy of generative models through which users access information, similar to search engines. To survive in this information economy might mean choosing proprietary systems, even if they preferred open models. Participants also felt more comfortable with the robust bias mitigations of proprietary models, as users stress test their conversational systems for (usually political) bias. Thus, for open models to present a viable alternative may require an ecosystem-level view [397, 469], accounting for not only model-related costs and internal usability, but also opportunity costs and external accessibility.

Finally, in the ninth chapter, I discuss the implications of the findings of the six studies presented in this work, and I describe the research directions I expect to pursue in the future, building on the present work.

Chapter 2
RELATED WORK

This section reviews the Related Work on which this dissertation builds, considering first the definitions of generative AI, general-purpose AI, and epistemic risk, before discussing existing manifestations and means of mitigating epistemic risk in these technologies, and presenting study-specific related work needed to meaningfully interpret the findings.

2.1 Generative and General-Purpose AI

Defining generative and general-purpose AI is challenging due to the many domains in which these technologies are now used. In the sections below, I provide precise definitions adequate to the purposes of the studies comprising this dissertation.

2.1.1 Generative AI

Generative AI can be technically described as "a type of machine learning architecture that uses AI algorithms to create novel data instances, drawing upon the patterns and relationships observed in the training data" [133]. The capacity to produce novel data instances distinguishes generative models from other machine learning methods and underlies the shift in how most users interact with most generative models: by "prompting" them [58, 422], rather than by programming them or further training them. Prompting a model refers to providing a series of example inputs and outputs such that, given a new input, the model generates an appropriate output (*i.e.,* a novel data instance conditioned on the input) [324, 58]. Consider a simple application that produces the phylum of an animal given its species. Given a set of examples like *tiger: chordate, beetle: arthropod, earthworm: annelid, frog:*, a generative model can produce *chordate*, and if allowed to continue generating, it

might produce a longer output string such as *chordate, cicada: arthropod, shark: chordate*. Generative language models can thus be prompted not only to capably sort the new animal into a phylum, but they can also generate new pairs of "species : phylum" data instances, despite not explicitly training to perform any task other than generating the next word in a sequence.

Generative models have been learned for many other modalities in addition to language. Generative image models can produce full images given a sequence of pixels [79]; generative models of music can produce songs given an audio sample [111]; and generative models of code can produce source code given a few lines of code or a comment [80]. Due to its benefits for usability, many generative models train to process multiple modalities [321], typically by using a language component as the primary means of interacting with the user. Generative text-to-image models, for example, allow the user to input text and receive a generated image as output [345], while generative speech-to-text models produce text conditioned on spoken language [322].

Recent advances in generative models simplify the process of prompting a model by training the model to infer the user's intent from a simple instruction. For example, rather than prompting the model with a set of species : phylum pairs, the user might instead provide the model an instruction like "Please output the phylum of the following animal: hawk", and the model would respond with a natural language answer like "chordate." Further extending the capability to generate novel data instances, instruction-tuned generative models can also respond to a request to produce data without conditioning on specific examples in the prompt; for example, the model could respond with a list of species: phylum pairs given the user prompt "Please output the phylum of any ten animals in a structured format such as species: phylum." This process of adapting a pretrained generative model to follow instructions is called *instruction-tuning* [231]. Though an instruction-tuned model can perform many tasks without receiving examples of these tasks from the user [421], providing examples still helps to carry out complicated instructions, and evaluations of popular instruction-tuned models often utilize many examples provided in-context [336, 280]. The now-ubiquitous chatbot format of

models like ChatGPT is a specific form of instruction tuning, wherein the model learns to respond to special tokens demarcating the message of the user and the model [398, 281].

Instruction-tuning typically precedes or occurs simultaneously with a method for aligning a generative model to human preferences, such as reinforcement learning from human feedback or direct preference optimization [290, 326], such that the data produced by generative models is not considered objectionable by end users. These methods have achieved significant reductions in explicit bias and toxicity in generative models, rendering interaction more palatable for most users [399].

2.1.2 General-Purpose AI

This dissertation will adopt the definition of Gutierrez et al. [161], who characterize general-purpose AI as "[a]n AI system that can accomplish or be adapted to accomplish a range of distinct tasks, including some for which it was not intentionally and specifically trained." For example, chatbot models like ChatGPT [281] might be used to summarize a scientific article, classify a review as positive or negative, or produce a draft of an email response to a colleague. To do so, the model would require no additional parameter updates, and using it for one task would not impede its ability to carry out the other tasks.

The characteristics of general-purpose AI bear some resemblance to those of "foundation models", a term coined by Bommasani et al. [51] to describe AI models that are "trained on broad data at scale and are adaptable to a wide range of downstream tasks." However, general-purpose AI differs in that foundation models are "central yet incomplete" [51]—that is, an application must be built on top of the base provided by the pretrained foundation model, generally by fine-tuning the foundation model for a task of interest. A fine-tuned foundation model might only be capable of performing a single specific task, even though the base foundation model can be adapted for many different such tasks. Conversely, one expects a general-purpose model to be usable for many tasks out-of-the-box; one does not think of ChatGPT, for example, as an incomplete model on which to construct something of use, but as a system capable of performing many tasks without additional training, even if it may also

be further fine-tuned to perform additional tasks more capably [283].

Whether a model or system constitutes general-purpose AI is more a question of degree than kind, as described by Gutierrez et al. [161], who note that a system can be made "more" general-purpose by accomplishing more distinct tasks. Though generative chatbot interfaces capable of inferring the user's intent presently permit perhaps the widest range of generality in responses to user input, zero-shot image associators like the CLIP language-and-image model [321] can be considered a more limited version of general-purpose AI. With no parameter updates and solely natural-language interaction, CLIP can be used in numerous computer vision settings, including image classification [321], retrieval [379], and scoring [170, 416]. Models like this expand the range of machine learning tasks possible without customizing the model, even if they do not seek to infer user intent, as facilitated by a chatbot interface. Similarly, future models will likely ingest and output data in many more domains and modalities than present-day chatbots, given attempts to expand user input channels even in present-day models like GPT-4o [284], meaning that the current state-of-the-art may someday seem like a relatively limited version of general-purpose AI.

Finally, note that general-purpose AI is also often described as a "general-purpose technology" [127], a term from economics meaning that its impact on productivity will affect every sector of the economy, rather than only one [60]. Unless otherwise stated, this dissertation refers not to the economic aspects of general-purpose AI but to the technical aspects described above.

*2.1.3 Why This Dissertation Studies Generative **and** General-Purpose AI*

Most of the models studied in this dissertation can be characterized as *both* generative and general-purpose. Instruction-tuned language models in particular possess the defining characteristics of both technologies. Moreover, even those models that cannot be described as generative are often employed as components in generative models; for example, CLIP models are often used in the pipelines of generative text-to-image models [329, 306].

Why, then, does this dissertation distinguish between generative AI and general-purpose

AI? The reason is that this permits the adoption of two distinct viewpoints on epistemic risk in these models. To study epistemic risk in a generative model is to study the risk in a technology capable of producing novel data, while to study epistemic risk in a general-purpose model is to study risk in a technology that is appropriate for use in many distinct tasks. Focusing on only one viewpoint would fail to account for the full impact of these models. Considering ChatGPT only as a model that can perform many different tasks ignores its capacity to *create* new data, like blogs or news stories. Considering CLIP solely in relation to its use in generative text-to-image models ignores its use in numerous other impactful settings, like zero-shot classification and image retrieval. This dissertation thus attends to epistemic risk as it concerns both generative and general-purpose AI.

2.1.4 Description of the AI Technologies Studied in This Dissertation

This section provides an overview of the forms of AI studied in chapters 3, 4, and 7 of this dissertation. Chapters 5, 6, and 8 consider sociotechnical approaches to generative and general-purpose AI more generally, rather than focusing on specific instantiations of the technologies. Note that I also provide specific detail about the technologies discussed below in the chapters that use or evaluate them.

Static Word Embeddings

Chapter 3 of this dissertation uses static word embeddings to study AI representations of teenagers. Static word embeddings are trained using deep neural networks to represent words as vectors based on the conditional probability of their co-occurrence with surrounding words [244, 91]. This dissertation studies FastText [49], an extension of Word2Vec [245] that incorporates information from subwords, and Global Vectors for Word Representation (GloVe) [302], which incorporates corpus-level statistics to improve the semantics of a representation. Word embeddings are now widely used in social science [36, 158] to study societal attitudes [142], because the cosine distance between word vectors captures information about semantic similarity [172]. Though static word embeddings can be used in a wide variety of NLP

settings, including as input to generative models like recursive neural networks [467], their use requires too much additional engineering effort for them to be considered general-purpose AI.

Generative Language Models

Chapters 3 and 7 of this dissertation study generative language models. Generative language models, sometimes referred to as "GPTs" (for "**G**enerative **P**retrained **T**ransformer") use a modified transformer deep learning architecture [407], employing solely decoder layers to generate an output conditioned on the preceding input [323]. GPT language models [324, 58, 281] are pretrained on the "causal" language modeling objective, taking as input a series of subword tokens and generating the predicted next token [98]. To provide a more naturalistic and reliable way of interacting for lay users, most modern GPT models are fine-tuned to adhere to the natural language instructions of a human user [290], producing a "chat" based interface wherein the model and a human user take turns, with the user typically providing an instruction or request [335]. Such models are typically trained to be safe and helpful to the end user via reinforcement learning from human feedback (RLHF) [20], or direct optimization using a language model as the reward model [326].

Chapter 3 of this dissertation studies three generative language models which are *not* fine-tuned as chatbots, as this study intends to observe differences between the human response to a freeform response and the model's response to the same prompt. These models include GPT-2 [324], LLaMA-2-7B [399], and distilGPT-2-Nepali. Chapter 7 of this dissertation studies five generative language models instruction-tuned to interact with the user as a chatbot, as this chapter is concerned with producing an approachable, chat-driven alternative to ChatGPT. These models include LLaMA-2-Chat-7B [399], Mistral-7B-Instruct [191], Falcon-7B-Instruct [7], OpenAI GPT-3.5-Turbo, and OpenAI GPT-4-Turbo [281, 288].

Note on Low-Resource Languages Chapter 3 of this dissertation studies Nepali, a "low-resource" language for which much less text data exists than other languages [32].

Much literature has established that the performance of models trained on low-resource languages is likely to lag behind that of higher-resource languages such as English [332]. While a multilingual model may improve performance in a low-resource language [355], its representations may also take on semantic properties and biases of a higher-resource languages (*e.g.*, English) [476, 331]. Because chapter 3 is concerned with detecting potential *differences* in English and Nepali, I use solely monolingual language modeling techniques, regardless of performance disparities, to ensure that I accurately capture the semantic properties of the target language, rather than the influence of a higher-resource language.

CLIP and Vision-Language AI

Chapter 4 of this dissertation studies facial impression biases in CLIP, a multimodal vision-language model pretrained using a symmetric cross-entropy loss [279, 474] to pair images with associated text captions [321]. After pretraining, CLIP can rank, retrieve, or classify images based on their association with text classes specified by a user at inference rather than pre-selected at the time of training, making it a "zero-shot" vision-language model [321], as well as a good source for semantically rich embeddings [435]. CLIP is composed of a language model (usually GPT-2 [324]), and an image encoder, such as a Vision Transformer ("ViT") [123] or a ResNet [167]. The language and image models are jointly pretrained, and representations are projected into a multimodal embedding space, in which cosine similarity quantifies the similarity between image and text [321]. In addition to standard CLIP models, chapter 4 studies "FaceCLIP" models trained by Zheng et al. (2022) [482], who introduce Facial Representation Learning (FaRL), which combines CLIP training with a masked image modeling objective [451] and trains on a faces-only subset of the LAION-400M dataset [359].

Impact of Scale in Deep Learning and in CLIP Chapter 4 specifically considers the impact of scale (expressed in pretraining dataset size and model parameters) on the extent to which CLIP learns facial impression bias. Prior research shows that the impact of dataset scale on deep learning models is empirically predictable [171] and that task performance

scales with training dataset size [380, 58]. Zhai et al. (2022) [468] empirically demonstrate that both model and data scale impact visual task performance, and set new state of the art on Imagenet [103] by efficiently scaling a ViT. In CLIP models, Cherti et al. (2022) [84] demonstrate a relationship between pretraining data scale and task performance. Prior work on AI bias also demonstrates increases in hate speech in CLIP models trained on larger uncurated datasets [42].

Text-to-Image Generators

Chapter 4 of this dissertation considers facial impression bias in Stable Diffusion text-to-image generators [345]. Most modern image generators are created using CLIP models like those discussed above, and one of the first uses of a CLIP model was to provide training supervision to OpenAI's first DALL-E image generator model [330]. Other text-to-image generators like VQGAN-CLIP similarly use CLIP embedding space measurements in their objective function [97]. Recent image generators such as Stable Diffusion 2 employ CLIP models as text encoders [345], passing CLIP text embeddings to a U-Net or similar latent diffusion architecture capable of generating an image conditioned on those text embeddings. More recently, DALL-E 3 ("unCLIP") decodes images directly from a CLIP embedding space, translating CLIP text embeddings into image embeddings, and inverting them [329].

2.2 Epistemic Risk

The contributions of this dissertation address *epistemic risk* in generative and general-purpose AI. I adopt the definition provided by Biddle et al. (2017) [39], who describe epistemic risk as "exposure to harm (in the broadest sense) from acting in the face of uncertainty." Biddle et al. intentionally deploy the term *epistemic risk* to encompass manifestations of risk in the process of producing knowledge that are not covered by the more narrowly scoped term *inductive risk*, which refers specifically to exposure to harm due to correctly or incorrectly accepting a hypothesis on the basis of given empirical evidence. In subsequent work, Biddle (2022) [38] considers the sources of epistemic risk in machine learning, specifically in the

context of a recidivism prediction system. He identifies six points of decision in the process of creating a machine learning system that involve "epistemic risk judgments", including: "(1) problem identification and framing, (2) data decisions and model competencies, (3) algorithm design: accuracy and explainability, (4) algorithm design: conceptions of fairness, (5) algorithm design: choices of outputs, and (6) deployment decisions about transparency and opacity" [38]. This description demonstrates the broad scope of epistemic risk, which can be assessed at many points in a a sociotechnical pipeline. It also bridges an important divide for this dissertation, taking a problem typically seen through the lens of algorithmic fairness or bias—machine-learned racial bias in a consequential social setting—and considering it through the lens of epistemic risk, or the potential harm done due to being wrong by virtue of flaws in the production of knowledge.

The epistemic risk posed by a generative or general-purpose model differs in meaningful ways from that of a narrowly scoped classifier. In their consideration of risk (not solely epistemic) presented by general-purpose AI models, Zanotti et al. (2024) [465] employ a multi-component model to characterize risk, breaking it down into *hazard*, or the source of harm (including its probability and magnitude); *exposure*, or who or what stands to be harmed; and *vulnerability*, or the level of susceptibility to harm. What makes risk in general-purpose AI distinct from risk in narrowly scoped machine learning models is its expansion of all three of these components. It expands hazard by introducing numerous situations in which a model could now be deployed, including many uses for which it was not explicitly intended; it expands exposure by significantly increasing the number of users capable of interacting with it via an approachable chat interface; and it increases vulnerability in that the authoritative and seemingly objective interface of such systems renders it more difficult for users to identify its shortcomings [465]. Envisioning how these considerations are relevant to epistemic risk rather than risk generally requires little additional work, as one need only restrict the range of tasks to those that produce information on which a user might rely.

The framework above offers further useful insight for the structure of this dissertation. As

noted by Zanotti et al. (2024) [465], hazard cannot be mitigated without first specifying the hazard, a sometimes challenging task for "multi-hazard" general-purpose models. Chapters 3 and 4 of this dissertation intend seek to accomplish this, addressing epistemic risk related to misrepresentations of individuals and groups by generative and general-purpose models. These misrepresentations can have harmful downstream effects: if a model learns subtle facial impression biases as described in chapter 4, a hiring decision might be made on the basis that the AI-driven system judged a person to be untrustworthy due to their facial features. Alternatively, the same model might be used in a dating app as part of an algorithm to match users based on personality traits inferred from their facial features. Thus, the contributions of the two studies primarily focused on characterizing epistemic risk largely involve the specification of hazard in generative and general-purpose models.

Zanotti et al. also address the mitigation of risk, noting that mitigation can include controlling the conditions under which people interact with a model (for example, implementing age restrictions), intervening at the "design level" (for example, constraining a model's capability to respond where it might produce undesirable output), and reducing the vulnerability of a population (potentially using automated tools, such as a spam filter in the case of email), as well as reducing the hazard itself, though this is not always possible [465]. This contributions of the this dissertation focusing on design and transparency can be considered design-level interventions and vulnerability-reduction interventions, in that they 1) produce strategies for using or positioning models in sociotechnical infrastructures that mitigate epistemic risk, as in chapters 5, 6, and 8; and 2) employ alternative technologies that reduce specific epistemic risks, such as those posed when using proprietary models, as in chapter 7.

2.3 Epistemic Risk in Generative and General-Purpose AI

This section provides an overview of specific sources of epistemic risk identified in prior work on generative and general-purpose AI. Comprehensively characterizing such sources would require more space than the dissertation can afford, so this section intends instead to provide an overview of the most salient sources of risk for the dissertation.

2.3.1 Characterizing Epistemic Risk

In this section, I address several model-specific sources of epistemic risk in generative and general-purpose AI, including Secondary Factuality, Non-Semantic Variance, and Social Misrepresentation.

Secondary Factuality: Despite their positioning as epistemic artifacts capable of delivering information about the world to end users [169, 460], epistemic risk arises from the training objectives of generative and general-purpose models like ChatGPT, which do not optimize for the factuality of their output. During pretraining, generative language models train to predict the most probable next word [323, 324, 58, 398], and during fine-tuning, they optimize to accord more closely with user preferences [290, 399]. Where a model produces factual output, it does so as a result of being trained for these objectives. However, the preference for the most probable output renders generative models vulnerable to *hallucination*, a phenomenon wherein the model produces plausible but non-factual data instances [190]. Efforts to increase the factuality of model output have focused on measuring and improving the truthfulness [225] and abstention [427] properties of such models, such that they either 1) answer a user correctly based on knowledge encoded in their parameters; or 2) abstain from answering the question. Preference alignment strategies like RLHF also improve the factuality of model output because in most contexts, human users prefer factual outputs over misinforming outputs [399]. Another common method known as retrieval-augmented generation (RAG) supplies the model with pertinent information from an external database or from the open internet [216, 262], positioning the model more as a vehicle for delivering information rather than as both the source and disseminator of information.

Non-Semantic Variance: Epistemic risk also arises from inconsistency in model responses between semantically interchangeable inputs. Recent work by Sclar et al. (2023) [360] demonstrates that non-semantic formatting differences in prompts (such as spacing and capitalization) can cause the performance of generative models to vary on common NLP tasks by up to 76 percentage points. Drawing on the example discussed previously, this means

that a model might output a meaningfully different answer depending on whether it received prompt A, *tiger: chordate, beetle: arthropod, earthworm: annelid, frog:*, or prompt B, *Tiger: Chordate, Beetle: Arthropod, Earthworm: Annelid, Frog:*. This introduces epistemic risk when using such models for tasks that benefit from reproducibility, as in many scientific settings. Underlining the potential susceptibility of research to prompt-based variance, reviews of recent social science research show that studies using language models often employ only one prompt, and that similar studies exhibit greater variance among reported model performance than might be expected [274]. Strategies for addressing prompt-driven variance include quantifying the difference induced by prompts using metrics like FormatSpread [360], as well as calibrating language models to account for baseline variance [477].

Social Misrepresentation: Epistemic risk in generative and general-purpose AI can concern the representation of social groups, which may be misrepresented or too narrowly represented by a model [364]. Numerous studies have shown that models trained on internet-scale web scrapes learn biased, oversimplified representations of marginalized groups, which generative models incorporate into the data they produce and general-purpose models transfer into the tasks for which they are employed [260, 2, 276]. My own prior work on CLIP models demonstrates the presence of hypodescent, also known as the one-drop rule [433]; markedness, wherein one demographic serves as a reference category for a model [437]; xenophobia, wherein American identity is conflated with White ethnic identity [434]; and sexual objectification of women [446]. These biases exist not only the CLIP model itself but also in generative models built utilizing CLIP [446, 434]. Even preference-aligned models like ChatGPT continue to misrepresent social groups, often in subtle ways: Cheng et al. (2023) [82], for example, demonstrate that text generated by chatbot personas reflects a default reference category that accords most closely with a White persona, while Omiye et al. (2023) [276] find that models propagate inaccurate, race-based medical theories. Approaches to mitigating misrepresentations of social groups vary widely, and they include debiasing methods for both the embedding space and the alignment process [417, 449], as well as community-engaged AI, which measures misrepresentation and seeks to address it in coordination with the people

most likely to be affected [109].

2.3.2 Epistemic Risk in Interaction

In this section, I address sources of epistemic risk that occur when users interact with generative and general-purpose AI, including Sycophancy, Latent Persuasion, and Implicit Bias.

Sycophancy: Producing models that accord with user preferences has mixed effects for epistemic risk. On the one hand, alignment to user preferences produces models that adhere more closely to user instructions and are more likely to correct infer user intent, and as noted above, users generally prefer factually accurate models that do not fabricate output [290, 399]. The exception to this rule, however, is that users also tend to prefer models that *agree with them*, even when they are not correct. Sharma et al. (2023) [363] demonstrate that models aligned to be helpful prefer responses that reflect the user's beliefs back to them, rather than accurate responses, an outcome the authors refer to as *sycophancy*, and which has subsequently been demonstrated to also occur in aligned vision-language models [479]. Similarly, Laban et al. (2023) [209] further show that aligned models change their response to agree with users after being challenged, even when the initial answer was correct. Thus, even aligned models have objectives that can supersede the dissemination of factual information, introducing epistemic risk. Strategies to mitigate sycophancy include explicitly aligning language models for "honesty" [455], as well as training on synthetic data to improve the robustness of the model [423].

Latent Persuasion: Interaction with generative and general-purpose AI may also sway users' perspectives without their conscious awareness. Jakesch et al. (2023) [188] found that individuals who wrote with the assistance of an opinionated AI co-writer were more likely to adopt the views of the model, both in the written text itself and in a subsequent survey, an effect the authors refer to as *latent persuasion*. Padmakumar et al. (2023) [291] further demonstrate that co-writing with language models "increases the similarity between the writings of different authors and reduces the overall lexical and content diversity." Though

these shifts do not necessarily connote an improvement or diminishment of information integrity, it introduces the possibility that users' epistemic values and practices will change in ways they cannot fully explain when producing information in collaboration with AI. Moreover, popular models like ChatGPT have recently incorporated *Memory* features, such that their output can be personalized to a user based on notes maintained by the model about the user [287]. The possibility of latent persuasion and confirmation bias may be further amplified when interacting a model that maintains significant knowledge about the user.

Implicit Bias: Finally, although aligning to user preferences reduces explicit toxicity [399, 290], including blatant misrepresentations of social groups, alignment strategies do not consistently reduce implicit and task-specific bias, which users might be less attuned to given the absence of explicit bias. Bai et al. (2024) [19] demonstrate that an array of instruction-tuned language models exhibit numerous implicit social biases that correlate with the model's bias in decision-making settings, such as the implicit consideration of race when a model is prompted to make a decision about hiring. Moreover, Acerbi et al. (2023) [3] demonstrate that ChatGPT directs attention to the most bias-congruent information in a text when prompted to recursively summarize a text, known as a "chain-of-transmission" bias in psychology. For example, a model might include information about a woman's role as a mother in a summary while leaving out information about her career, and the issue will become more magnified as the chain of summarizations continues. Implicit bias in the model bears some similarities to latent persuasion, but differs in that the user is not persuaded but tacitly encouraged to accept an epistemically flawed outcome influenced by societal bias.

2.3.3 Epistemic Risk in Proprietary and Open Models

In this section, I address sources of epistemic risk that occur when users interact with proprietary generative and general-purpose models, including Uncertain Data Provenance, Hidden Processes, and Observer Effects. This section requires a definition of open and proprietary models, for which I draw on the work of Palmer et al. (2024) [293], who characterize open models as those that can be downloaded, run outside of an API, and

versioned by the user, and for which the model's training data is fully disclosed, even if the data itself is not made accessible. Conversely, proprietary models cannot be downloaded or versioned by the user, and they may thus be updated by the external organization serving the model, potentially without notifying the user [275]. They cannot be run outside of an API, and their training data is usually not disclosed [280, 275]. As discussed below, proprietary models introduce epistemic risk in two ways: by limiting what the user knows about the model and system they are using, and by introducing the possibility that the provider will *change* what the user knows about the system they are using, potentially without informing the user.

Uncertain Data Provenance: As noted by Biddle (2022) [38], the choice of training data is a source of epistemic risk for a machine learning model. However, in the case of proprietary models, users do not know what data the model was trained on at any stage in the production of the model, ranging from pretraining to alignment to additional fine-tuning [280, 293]. Though characteristic of proprietary models, the non-disclosure of training data sources also affects some "open-weight" models such as Mistral-7B [191], a model studied in dissertation. Model developers often discuss the confidentiality of a model's data sources as key to competitive advantage, though such opaqueness necessarily increases uncertainty about a model's fitness for a task, and thus also increases epistemic risk. Though attempts at more comprehensive benchmarks like Stanford's Holistic Evaluation of Language Models (HELM) [223] can help to mitigate some of this risk by measuring model performance across many domains, they do not serve as a substitute for disclosed data sources, which can themselves be studied for insight into the risks of training on them [45].

Hidden Processes: Using proprietary models increases epistemic risk because the processes by which such models produce their output are hidden from the user. Consider that OpenAI's o1 model, released in September 2024, intentionally hid the raw chain-of-thought reasoning process used to produce its output [286]. Rather than providing this output to the user, who might gain valuable information about whether the model's reasoning process was sound, OpenAI makes only a model-generated summary of that reasoning available to the user [286].

Model providers usually place restrictions on what users see about their models in order to protect them from reverse engineering approaches [72]. However, hiding the true reasoning process of a reasoning model, and suppressing the full probability distribution used to produce linguistic output, can increase uncertainty about the validity of how a model arrived at its output, and thus also increase epistemic risk.

Observer Effects: Finally, corporate entities often update their models to prevent undesirable behaviors, without providing access to the previous model [275], which may have been used to obtain results of epistemic import. This may be socially desirable, but it introduces what Holtzman et al. (2023) [177] characterize as "observer effects" in scientific studies, wherein research on observing an undesirable behavior in the model is no longer reproducible *by virtue* of being observed. One notable example of this includes OpenAI's attempt to mitigate the lack of demographic diversity in DALL-E 2 outputs by postprocessing user prompts without the user's knowledge [282, 269]; for example, the prompt "a portrait photo of a firefighter" might be suffixed with "who is female and Asian." Like the hidden processes discussed above, the epistemic risk of observer effects concerns uncertainty about how a model arrived at its output, but observer effects also consider changes in the visibility of epistemic processes that that occur due to observed model behavior. Open models can mitigate observer effects, because they are downloadable, and previous model checkpoints typically remain available [293], albeit with developer warnings.

2.4 Study-Specific Background

While the studies included in this dissertation are unified in presenting new approaches to epistemic risk, they are also informed by a diverse variety of fields, including information science, computer science, and psychology and the social sciences. The purpose of this section is thus to provide an overview of the background literature relevant to each of the studies included in this dissertation.

2.4.1 Teenagers and Representation Bias

Chapter 3 of this dissertation considers the epistemic risk of systematically misrepresenting a social group—in this case, teenagers (or adolescents). In this section I review the Related Work on societal representations of teenagers, as well as on age-related biases in AI.

Defining Adolescence

The National Institutes of Health (NIH) define Adolescents as persons between 13 and 17 years old, distinct from Children (1 through 12), Adults (18 and older), and Older Adults (65 and older) [265]. While definitions may vary between cultures and across time [13], in this dissertation I adopt the NIH definition, which is consistent with the related work below.

Media Representations of Adolescents

Prior work finds that popular and news media depictions of adolescents are generally negative, with positive interactions involving teenagers portrayed as deviations from the norm [31]. News coverage of teenagers often depicts supposed epidemics of violence, crime, drug abuse, mental illness, and immorality, which are usually not well supported by evidence [148, 389]. In foundational work, Dorfman et al. (1997) [119] find that most California TV news reports related to violence feature youth, and that only education policy receives as much treatment as violence in newspaper coverage about adolescents. Males (1999) [237] find that LA Times articles included adolescents in stories about violence five times more frequently than adults. Adolescent behavior may be presented as dangerous even when not volitional, as Best (2008) [34] find that activities as simple as teenage driving can be framed as pressing issues in the media. More recently, teenage use of technology has become a subject of public concern, and Stern and Burke Odland (2017) [373] find that print and online news media portray teens as having an unhealthy relationship with social media. Previously, Stern (2005) [372] found that U.S. films depict teenagers as violent, self-absorbed, and disengaged from civic life. In chapter 3, I draw specifically on the work of Stern (2005) [372] and Stern and Burke Odland

(2017) [373] to create prompts for evaluating the output of generative language models.

Societal Impact

Media depictions shape adult views of adolescents and may shape adolescent behavior. Hancock (2001) [164] shows that adults overestimate and perceive illusory increases in adolescent crime. Aubrun and Grady (2000) [17] find most adults report good experiences with teenagers they know but consider such experiences atypical, rather than questioning media framing. Dorfman and Schiraldi (2001) [118] note that negative media portrayals, especially of adolescents of color, lend justification to harsher treatment and more restrictive policies. Moreover, Qu et al. (2020) [318] find that younger teens' *own* beliefs in teenage stereotypes contribute to behavioral problems. Buchanan et al. (2023) [62] argue that, to prevent a self-fulfilling prophecy, commonplace descriptions of adolescent "stress and storm" must be replaced with a less reductive framing, such as "possibility and promise."

Societal Variation

Though some aspects of adolescence appear consistent around the world [371], scholars describe significant variation in characterizations of adolescence both within and across cultures [63]. Enright et al. (1987) [128] note that definitions of adolescence change over time based on society's needs: during war time, teens are portrayed as rugged and adultlike, but when they are not desired in the workforce, teens are portrayed as more childlike. Arnett (1999) [13] note that adolescent stress may be more pronounced in individualistic western cultures, while Larson and Wilson (2004) [212] use the plural form "adolescences" to describe variations around the world and across time, noting that teen years are not consistently characterized by emotional turmoil and psychic separation from parents. Finally, Di Giunta et al. (2023) [112] observe differences in emotion regulation in teenagers in Italy and Colombia, suggesting cultural factors play a role in adolescent well-being. The study discussed in chapter 3 of this dissertation considers the outputs of both English and Nepali AI models and the

perspectives of both English and Nepali participants to provide insight into the potential variation of this representation bias across languages and societies.

Age Biases in AI

Considering the flawed representation of teenagers in AI also necessitates that I locate chapter 3 in the literature on age-related biases in AI. Prior research on age-related biases in AI describes the functional failure of technologies like emotion recognition for older adults [202], often precipitated by underrepresentation of older adults in training data [297]. Studies of young/old bias in static word embeddings find that youth is preferable to old age [68, 113, 383] but they do not analyze adolescents as a distinct age group. Studies of biases in multimodal language-vision models, on the other hand, have identified several undesirable biases associated with teenagers. Agarwal et al. (2021) [5] find that OpenAI's CLIP [321] associates criminality with images of adolescents. In my own prior work [446], I find that CLIP exhibits sexual objectification biases and that text-to-image generators like Stable Diffusion [345] output sexually objectifying images of teenage girls.

2.4.2 Facial Impression Bias

Chapter 4 of this dissertation considers the epistemic risk of AI making unwarranted inferences about human faces. In this section I review the Related Work on facial impression biases in human society, as well as computational approaches to studying facial impression biases.

Facial Impression Bias

A wealth of psychological research indicates that humans make immediate judgments about the attributes of people they do not know based solely on facial appearance [430, 271, 76]. Information inferred from faces includes character traits (like trustworthiness and outgoingness) and socially constructed group memberships (like gender and ethnicity), as well as relatively objective traits (like hair color and weight) [396, 304]. Research on facial first-impression

biases in humans has found that the inference of attributes from facial appearance plays a role in numerous consequential domains, including employment decisions [374, 151, 382], criminal sentencing [431, 194], and the election of political candidates [11, 273, 214, 185]. While facial impression biases may be consistent among a population, inferences of unobservable attributes such as character traits are inaccurate and often reflect societal stereotypes [381, 396]. AI systems are increasingly employed to automate or mediate access to information in domains such as hiring [222], political analysis and advertising [294], and law [87], and to the extent that such systems reflect facial impression biases, they may have socially undesirable impacts.

Relationship to Social Group Biases

Some studies suggest a connection between facial impression biases and biases related to demographic traits such as gender and ethnicity. Oh et al. (2019) [271] find that gender biases associating men with competence are reflected in participant impressions of the competence of faces. Xie et al. (2021) [450] find that the structure of impressions of novel faces is predicted by learned social stereotypes about gender and race. Peterson et al. (2022) [304] find that facial impression biases are correlated with demographic categories, such that judgments of traits like "cuteness" are related to age. The relationship between facial impression bias and social stereotypes can have real-world consequences. For example, prior work finds that White phenotypic prototypicality (looking like the average White person) can moderate use of force by police [196].

Computational Models of Facial Impression Bias

Chapter 4 evaluates the presence of facial impression biases in CLIP and Stable Diffusion as a manifestation of epistemic risk in a general-purpose AI system. However, much prior work intentionally creates computational models of facial impression biases to aid in scientific study. Most recently, Peterson et al. (2022) [304] collect facial impression ratings from human subjects and use them to create a model facial impressions using the StyleGAN-2 network [199], demonstrating its capacity to manipulate faces such that the average U.S. perceiver

would consider them similar to an attribute (such as trustworthiness). They build on research on the scientific modeling of facial impression biases, which commonly utilizes techniques including landmark annotations of faces [403], parametric three-dimensional mesh modeling [48], geometric morphological analysis [353], and supervised deep learning models [459]. As noted by Peterson et al. (2022) [304], creating a computational model of a *bias* differs from modeling the attribute itself (*i.e.,* trying to predict if an individual is trustworthy from their face, rather than whether the average person would *perceive* an individual as trustworthy), which would amount to physiognomy [453] for an unobservable attribute like trustworthiness. The reflection and dissemination of such unwarranted inferences is precisely the concern, though, with models like CLIP, which do not intentionally learn these biases for scientific study, and which may reflect them to lay users.

2.4.3 Fact-Checking and Sociotechnical Infrastructures

Chapter 5 of this dissertation studies epistemic risk in the context of fact-checking, specifically asking the question of how generative AI is affecting fact-checking before contributing a "dimension of verification" to the design space of generative models. This section reviews the related work on fact-checking, including its sociotechnical infrastructure, disciplinary values, and points of contact with generative AI.

Sociotechnical Infrastructures of Fact-Checking

"Fact-checking" refers to the investigation of potentially misinforming claims and narratives that may adversely impact individuals and society [154, 156]. While a version of fact-checking has long existed in the form of investigative journalism [114], modern fact-checking coincides with the rise of the internet and social media in particular [156], which provided new conduits for the spread of misinforming content among vast networks of people. Fact-checking is primarily a "socio-technical" task [88, 458, 325], wherein technology is useful and meaningful only in the context of its relationship to the humans who interact with it [385, 463]. While fact-checkers necessarily employ data-driven technologies [160, 115], and envision further

uses of technologies to, for example, minimize the amount of harmful content to which they are exposed [195], human judgment is also crucial to the fact-checking process [155], and fact-checkers are skeptical of technologies that promise to fully automate parts of fact-checking work [195]. Prior work has sought to clarify the communities [57] and sociotechnical infrastructure undergirding the processes of fact-checking. Juneja and Mitra (2022) [195] describes fact-checking organizations as composed of "human and algorithmic infrastructures" fulfilling distinct roles in fact-checking, such as editing and investigation. Chapter 5 of this dissertation builds on these roles to describe the opportunities presented by generative AI in fact-checking.

Prior work has discussed some of tradeoffs described by journalists and fact-checkers in adopting generative AI. Such technologies can both pose difficulties for fact-checkers, who must contend with higher quality misinformation produced more easily [462, 198, 187], but also opportunities for novel technologies for supporting their work [340, 100]. Recent work highlights difficulties with generative AI for journalism and fact-checking, including low audience trust in AI-generated content [230] and algorithmic biases in the dissemination of AI-assisted fact-checks [263].

Technology Adoption and Organizational Change

One of the chief concerns of Chapter 5 pertains to how generative AI will reshape the profession of fact-checking, and to that end we review the relevant related work on technology-driven organizational change. The primary framework used to describe organizational change in chapter 5 is the Technology-Organization-Environment (TOE) framework of Prasad Agrawal (2023) [313], which provides a means of breaking down factors motivating technology adoption based on elements like regulation (an environmental or societal concern) and organization size (an organizational concern). Though chapter 5 poses questions related to generative AI, its potential as a *general-purpose* technology capable of automating many tasks previously performed by humans looms large in the challenges and opportunities identified by fact-checkers. In the context of organizations, Brynjolfsson et al. (2021) [60] describes general

purpose technologies as necessitating intangible "complementary investments" to realize their potential, such as "co-invention of new processes, products, business models and human capital," suggesting the sociotechnical and potentially transformative nature of technology adoption. However, in a now foundational text, Fichman and Kemerer (1999) [134] also note that the widespread acquisition of a technology by organizations may not result in its widespread *deployment*, especially where "knowledge barriers" mitigate effective use. Chapter 5 provides a consideration both of what new processes and products might be enabled by generative AI, and what barriers might prevent its adoption or mitigate its ultimate effectiveness as an epistemic tool.

Human-Centered Design and the Values of Fact-Checking and AI

Researchers in HCI have mapped the design space [71] of generative AI [255], describing interactions possible with users and ways to use it in domains like scientific research [254] and creative writing [74]. Building on participatory design [47, 370], recent work develops "participatory AI," wherein human subjects envision new AI-driven designs with researchers [102, 43, 319]. In fact-checking, Das et al. (2023) [100] conduct a review of human-centered NLP and develop a confusion matrix for calibrating trust in human-AI collaborations.

Chapter 5 contributes to the design space of generative models by drawing on interviews with fact-checkers who are primarily members of the International Fact Checking Network (IFCN) [311]. These fact-checkers adhere to the IFCN's rigorous ethical codes [305], which are made publicly available and to which signatories must commit prior to admission in the IFCN [312]. The principles set forth by the IFCN and reflected in the interviews are useful both for proposing a socio-technical dimension of verification in the design space of generative models, and for considering the potential tensions between fact-checking and generative AI. For example, scholars have found that AI and machine learning research is not "value neutral" but prioritizes values like performance and generalization, while neglecting considerations like "negative potential" [44]. Thus, chapter 5 *also* uses the IFCN principles directly to chart value tensions between generative AI and fact-checking.

2.4.4 Nonviolent Communication and Needs-Conscious Design

Chapter 6 of the dissertation utilizes the principles of Nonviolent Communication (NVC) to create a design framework called Needs-Conscious Design, which prioritizes empowering people to meet their interpersonal needs using technology, including AI. This section reviews the related work on NVC, designing for interpersonal connection and well-being, and the use of AI to support these forms of design.

Nonviolent Communication

Nonviolent Communication, or NVC, is an approach to structured communication intended to meet human needs [346]. NVC employs a four-component approach to communication, characterized by 1) observing the situation without making judgments; 2) accurately naming one's own feelings; 3) linking feelings to the underlying needs from which they arise; and 4) making requests for those needs to be met [346]. While NVC grew out of Rogerian person-centered psychology [343, 344], it has seen much application in the realm of conflict mediation, as NVC was implemented to facilitate racial integration of American schools [346], and it has been used, sometimes alongside mindfulness training [376], in schools [347], in prisons [239, 376], and in healthcare settings [259, 266, 411].

Outside of isolated projects [137] built around its core tenants, NVC is not a common subject of study in design or in Human-Computer Interaction (HCI). However, its emphasis on structured communication and widely shared human needs exhibits commonalities with psychological theories commonly applied in HCI, such as Basic Psychological Needs Theory (BPNT) [406], a subtheory of Self-Determination Theory (SDT) [405]. SDT holds that individuals' well-being derives from three basic psychological needs: autonomy, competence, and relatedness [404]. Research in HCI sometimes evaluates the success of a technology based on the satisfaction of these needs [478].

Designing for Interpersonal Connection

Chapter 6 builds on prior work on designing to support human connection, including frameworks for meeting social needs; thus, I review prior work on such frameworks here. Baughan et al. (2021) [22] introduce Interpersonal Design, an approach that centers relationships in the design of technologies, and which has seen further development in the context of designing for conflict resolution and good faith disagreement [23]. Similarly, Zhang et al. (2021) [472] propose designing for emotional well-being, an approach that produces ways to nudge users into prosocial interactions likely to increase their well-being, so long as those interactions do not feel emotionally burdensome. Marcu and Huh-Yoo (2023) [238] introduce Attachment-Informed Design, a set of principles for supporting both relationships and communities in the context of mental health interventions. Shakeri et al. (2024) [361] design smart home technologies to support "passive co-presence" at a distance, allowing users to experience "passive aspects of family life" when they are not physically close. Kim et al. (2024) [204] identify features in the design of the BeReal social media platform that facilitate authentic self-presentation. Finally, Kelly et al. (2017) [201] introduce Effortful Communication, suggesting that designing for interpersonal relationships should support communication that is "effortful" or "demanding by design" and characterized by "discretionary investment, personal craft, focused time, responsiveness to the recipient, and challenge to a sender's capacities".

Designing to Support Well-Being

Because chapter 6 intends to produce designs that help users meet their emotional and interpersonal needs, it also builds on prior research focused on supporting user well-being. Hoefer and Voida (2023) [175] study the satisfaction of user needs in everyday situations, showing how the design of a personal informatics system might help users to reflect on and address common needs. Much work in designing for user well-being employs the construct of mindfulness. Li et al. (2023) [220] conducted workshops with mindfulness practitioners to design technologies to support mindfulness experiences for users of many experience levels.

Tan et al. (2023) [384] introduce Mindful Moments, a design that uses the affordances of smart-glasses to support well-being. Hsu et al. (2023) [180] center identity, connectedness, security, and autonomy to inform the co-design of social robots with people living with dementia. Arné (2024) [12] study the impact of digital technologies on user motivation toward self-reflection, drawing on SDT to inform their designs. Kim et al. (2024) [203] study how to reduce dysfunctional privacy concerns on social media to support interpersonal connection. Nonviolent Communication is well-known for helping practitioners meet needs via the precise use of language, and recent work in design also supports well-being by leveraging text-driven technologies, ranging from smartphone messaging to chatbot language models. Bhattacharjee et al. (2023) [37] develop an approach to contextual messaging to support user psychological well-being, adjusting characteristics of messages for users experiencing low mood. Park et al. (2021) [299] design a chatbot assistant to support expressive writing about mental health experiences, finding that the bot can "encourage narrative writing, with relative ease and emotional disclosure." Moreover, Fu et al. (2023) [140] design a conversational agent to help children learn positive self-talk to manage emotional distress.

AI in Interpersonal Design

Many of the participants in the study discussed in chapter 6 considered the role AI might play in empathetic designs, and much recent work examines the possibilities and challenges of using AI in designing to support interpersonal relationships. As described by Dö (2023) [120] in a study of fictional social robots, design work often conceptualizes chatbots as either *substituting* humans in relationships, or *mediating* human relationships. Work approaching AI as a *mediator* of human relationships includes that of Zhang et al. (2023) [470], who find that humanoid robots and computer screens can provide "conversation facilitators" that break the ice and engender deep conversations between strangers. Similarly, Shin et al. (2023) [366] find that a chatbot can facilitate online discussions, using the social media data of users to familiarize users with each other. Fu et al. (2023) [139] find that users prefer AI-mediated communication tools more in formal situations than in informal situations. Capel et al. (2024)

[70] explore the use of generative AI for self-care, including for social simulations, finding that text-based stories from AI allow some users to immerse themselves in a simulated social scenario.

Design work exploring AI as a *substitute* for certain human interactions includes that of Xygkou et al. (2023) [452], who describe seven scenarios in which chatbots can be employed to offer emotional support to mourners during a time of grief, including acting as a friend, listener, emotion coach, romantic partner, or simulation of the deceased. In some cases, AI may be used to stand in for a specific person, potentially someone known to the user. Lee et al. (2023) [213] find that believable "AI clones" of a person can cause "doppelganger-phobia" (negative reaction to displaced identity), "identity fragmentation" (threats to self-perception and individuality), and "living memories" (over-attachment to the AI clone of someone that one already knows). Chapter 6 surfaces additional concerns about AI-driven design and point instead toward design that the cautiously integrates AI, drawing on the principles of NVC to support relationships.

2.4.5 Open and Proprietary Models

Chapters 7 and 8 of this dissertation consider questions of transparency, first from a technical perspective and then from a sociotechnical perspective. In doing so, they discuss the implementation and tradeoffs of open and proprietary generative and general-purpose AI models. This section thus discusses what is meant by the terms open and proprietary in the context of this work. While scholars have observed a spectrum of relative openness in AI releases [368, 211], this dissertation adopts the definition of open models proposed by Palmer et al. (2024) [293] and Rogers et al. (2023) [342], as noted previously. That is, open models can be: 1) downloaded locally; 2) run without a call to an API; 3) shared with others; and 4) versioned. The contents of open models' training data must be disclosed, even if the data itself is not available [293]. This definition is satisfied by recent releases of generative language models such as Meta's LLaMA-2 [399] and its finetuned variants such as Stable Vicuna [85]. Some generative models, such as the popular Mistral-7B, are better characterized as "open

weight" models [191], as they make the model's weights available, but do not disclose training data in order to preserve competitive advantage. While many transformer-based models, such as Google's BERT [110], would qualify as open models under these definitions, chapters 7 and 8 of this dissertation are concerned specifically with *generative* models.

Scale As A Challenge to Transparent AI

Most generative models use billions of trainable parameters to mimic human language and learn to recognize other modalities [399, 191, 7]. Training or deploying such models requires vast financial resources, making them difficult to create and access for researchers and public interest practitioners [28]. However, while pretraining these models remains prohibitively expensive, recent techniques mitigate the difficulties of using models with low-cost hardware. For example, quantization loads a model in a lower level of precision than used during pretraining [106]. While most generative models are pretrained in 32-bit or mixed 32/16-bit precision, quantization loads the weights in 8-bit [106], 4-bit [108], or even 2-bit precision [108, 78], reducing memory demands. Because transformer models achieve much greater speed both during inference when using GPUs, the bottleneck for efficiently deploying a model is often the amount of memory (Video-RAM) on the GPU device [107]. However, quantization alone does not permit efficient *training* on commercial grade hardware [106]. Methods known as *parameter efficient fine-tuning* (PEFT) attempt to preserve the general-purpose functionality of a pretrained model while adapting it for a specific task [464, 117]. Among the most widely used PEFT techniques is low-rank adaptation (LoRA) [181]. LoRA inserts small, trainable weight matrices into a pretrained model, which are fine-tuned while leaving the learned parameters of the pretrained model unchanged [181], reducing fine-tuning memory costs. LoRA weights also require less space than a fully fine-tuned model [181]. Saving a fine-tuned LLaMA-2-7B-Chat model would require about 13.5GB of storage; saving only fine-tuned LoRA weights—which can later be inserted into the pretrained model—requires only around 260MB [181, 107]. Dettmers et al. (2023) [107] introduced qLoRA, a method for allowing trainable LoRA weights to be inserted into quantized models, allowing relatively

large models to be mounted on a small GPU and customized using LoRA [107]. Though the quantization and low-rank adaptation are important for any form of personalized AI that involves changing a model's weights, chapter 7 of this dissertation specifically uses qLoRA to assess the viability of small, adapted models as an alternative to large, closed models like ChatGPT.

Benchmarking and Evaluating Small-Scale Models Against ChatGPT

The popularity of ChatGPT has rendered it a reference point for researchers evaluating traditional NLP methods and models against the latest in generative AI. Kocoń (2023) [206] evaluate ChatGPT against state-of-the-art NLP models on 25 specific NLP tasks, finding that GPT-3.5-Turbo and GPT-4 are outperformed by these models and methods. Thalken et al. (2023) [391] show that a fine-tuned LEGAL-BERT [75] is the best-performing model for classifying legal reasoning, outperforming models like GPT-4 and LLaMA-2-Chat. Loukas et al. (2023) [232] find that fine-tuned sentence transformer models outperform few-shot GPT-3.5-Turbo and GPT-4 on a financial text classification task. Wang et al. (2023) [419] find that a fine-tuned BERT can outperform ChatGPT on sentiment analysis. Chapter 7 builds on this prior work by focusing specifically on adapting small, generative, chat-based models—which are more approachable to non-expert users—to be more competitive with ChatGPT through fine-tuning.

2.4.6 Human-Centered Data Science

Chapter 8 of this dissertation focuses on the sociotechnical considerations surrounding the adoption of open models in organizations, focusing on fact-checking organizations specifically. To do this, chapter 8 adopts a human-centered data science approach that locates uses of open vs. proprietary models in specific *data pipelines* employed by human fact-checkers. Note that this approach is more specific and more concerned with real-world data practices than the organizational infrastructures discussed in chapter 5. This section thus discusses the related work on human-centered data science.

As noted by Berman (2023) [30], interaction between data practitioners and the tools they use constitutes a social context that shapes the ethics of AI practices in organizations. Much work in social computing seeks to describe these tools and their interaction with human practitioners in data science *pipelines* [415]. For example, in a study of 183 data scientists, Zhang et al. (2020) [471] describe a common data science pipeline consisting of three high-level steps, including preparation, modeling, and deployment. However, as described by Hopkins and Booth (2021) [178], studies of data practitioners typically center on participants from big tech and academia, and may overlook the challenges faced by smaller organizations facing resource constraints, such as tensions between user privacy and organizational growth, a finding also echoed in Bessen et al. (2022) [33], who note that AI startups may face tradeoffs between building more competitive and more ethical products. Human-centered data science centers the context in which data practitioners perform their work, acknowledging that data work work may be undertaken by domain experts or other workers not traditionally considered data scientists [258], a perspective that can yield domain-specific understandings of data pipelines. For example, Rothschild et al. (2022) [348] note that civic workers at public and non-profit institutions exhibit skill with data contextualization that provides value far in excess of their sometimes less-developed computational abilities.

Adopting a human-centered approach in chapter 8 is well-motivated for fact checking, because despite the many benchmarks and techniques to support detection of misinformation [350, 327, 86], fact-checking organizations often view academic research as too detached from the real world [195], and recent work argues that even core NLP research on fact-checking should also study human factors [100]. Chapter 8 thus privileges the views of fact-checking professionals, surfacing where generative models fit into fact-checking data pipelines, and contextualizing the value fact-checking organizations see in open and proprietary models within those pipelines.

Chapter 3

THE RISK OF MISREPRESENTATION: A BILINGUAL, BICULTURAL STUDY OF ADOLESCENT REPRESENTATION BIAS IN AI

3.1 Preface

Research on bias in language models often focuses on a form of bias that people widely agree is unacceptable, such as gender bias in models used for job screening [9], or skin tone bias in image classifiers in facial recognition systems [328]. Yet some demographic biases remain common enough that they are routinely repeated in the popular and news media. In the first study of my dissertation, I will demonstrate the epistemic risk induced by training general-purpose models on high-quality media data out of context: namely, a disconnect between the representations learned by these models and the perceptions and experiences of vulnerable and marginalized people thus represented. The findings of this chapter support **Thesis Statement 1**: *Epistemic risk in generative and general-purpose AI models results not only from explicitly toxic training data but from using ethical, high-quality data sources outside of their original context in order to train AI.*

3.2 Introduction

Teenagers feature more prominently in western media accounts of new technologies than perhaps any other user group. They are the group most likely to adopt and capably use new technologies, including social media [409] and ChatGPT [205]. However, to read media accounts, they are also the most likely to *misuse* new technologies, leading to harm to others, or inadvertent harm to themselves [373]. Such narratives have consequences for adolescent access to technology: concerns about compulsive use of social media, cyberbullying, and

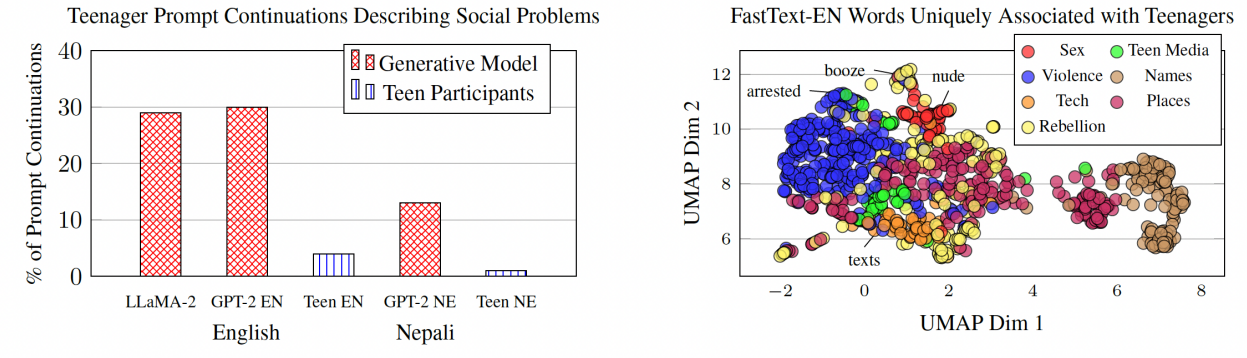


Figure 3.1: Left: Teenage participants were much less likely to continue prompts about teenagers with social problems than generative language models. Right: Words associated with adolescents over any other age in English static word embeddings reflect violence, rebellion, and sexualization.

sexual predation led to a March 2024 ban on use of numerous social media platforms by younger teenagers in the state of Florida [392]. Concerns about deceptive design and online safety warrant consideration; yet the response—a blanket ban—suggests a framing that emphasizes the danger of adolescent technology use and affords adolescents little agency.

Such presentations continue a decades-long trend in western media portraying teenagers as simultaneously *a risk* to society and *at risk* from society [292]. Though largely disconnected from most adults' experiences with teens [17], media portrayals of adolescents have centered violence, drug abuse, sexualization, technology addiction, and even religious fanaticism as pressing issues that warrant responses ranging from targeted media campaigns to government legislation [89, 240, 148, 389]. Though such portrayals seem sensationalistic in hindsight, representations of teenagers in media sources nonetheless shape adults' beliefs about what adolescents are like, influencing the treatment of adolescents in public places [31] and the restrictiveness of policy intended to influence adolescent behavior [118].

In the present work, I study societal attitudes toward adolescents learned by static word embeddings and generative language models, comparing with attitudes reported by

adolescents themselves. Because prior work suggests attitudes toward adolescents vary across cultures [212, 112], I undertake a bilingual, bicultural study, examining U.S. attitudes and English-language models, as well as models trained on Nepali, a low-resource language spoken primarily in Nepal, a South Asian country in the Global South, and a native language for my first co-author on this research. I held workshops with N=13 English-speaking adolescents in the U.S. and N=18 Nepali-speaking adolescents in Nepal, asking how adolescents *are* represented in media, and how they *should* be represented in AI. I make three contributions:

- **I show that English-language static word embeddings and generative language models associate adolescents predominantly with social problems.** Clustering the 1,000 words most associated with teenagers in English GloVe and FastText static word embeddings reveals that clusters related to drugs, rebellion, violence, mental illness, stereotypes, and sexual taboo account for more than 50% of words in GloVe and more than 40% in FastText. Similarly, using prompts about teenagers derived from Stern (2005) [372], I show that 29% of English LLaMA-2-7B outputs and 30% of GPT2-XL outputs depict societal problems. Of these, 47% depict violence in LLaMA-2, and 50% depict violence in GPT2-XL. Many such outputs mimic the format of "high-quality" training data—newspapers and journalistic media. Only 13% of distilGPT2 Nepali continuations reflect societal problems, and 10.1% of the words most associated with teenager in Nepali GloVe describe societal problems.

- **I show that AI representations are disconnected from adolescent self-perceptions.** Adolescent ratings of their own traits are decorrelated from static word embedding associations between corresponding trait vectors and the *teenager* vector, with Pearson's ρ=.02, *n.s.* in FastText and ρ=.06, *n.s.* in GloVe for English; and ρ=.06, *n.s.* in FastText and ρ=−.23, *n.s.* in GloVe for Nepali. Participant continuations of the same prompts used with generative language models show social problems arise in fewer than 4% of U.S. teen continuations and fewer than 1% of Nepalese teen continuations.

- **I discuss two central concerns of participants for fair representation in AI: diversity and positivity.** U.S. and Nepalese participants were aware of adolescent media stereotypes, and noted the difficulty in achieving fair representation. U.S. participants stressed that AI should foreground the *diversity* of teenagers, while Nepalese participants stressed that AI should present the positive traits of teenagers. Both groups expressed optimism that AI could correct media stereotypes about adolescents.

This work shows that generative language models learn societal biases latent in media framings. As user-facing models are integrated into schools and other contexts where they will impact adolescents' lives, research must center participatory approaches to AI [102] to ensure groups with less agency, like adolescents, are represented in ways that capture not a media presentation but a group's understanding of itself.

3.3 Models and Training Data

The present work studies monolingual static word embeddings and generative language models in English and in Nepali. I examine the following **static word embeddings**:

- **GloVe-CC**, 300-dimensional (300d) English-language GloVe embeddings pretrained by Pennington et al. (2014) [302] on the 840-billion token Common Crawl circa 2014 [95].

- **FastText-CC**, 300d FastText embeddings pretrained by Bojanowski et al. (2017) [49] on a filtered and deduplicated version of Common Crawl.

- **GloVe-NE**, 300d GloVe embeddings that I trained for the study, discussed further below.

- **FastText-NE**, 300d FastText embeddings pretrained by Grave (2018) [153] on Nepali Wikipedia.

FastText embeddings like FastText-NE are among the most used low-resource models for social science [226]. I trained a Nepali GloVe embedding after considering several pretrained

Nepali embeddings, including the NPVec1 model of Koirala et al. (2021) [208], the Nepali Word2Vec model of Lamsal (2019) [210], and the model of Subedi and Poudyal (2023) [378]. I ultimately trained an embedding on the dataset of Timilsina et al. (2022) [394] because it contained three times the data (800 million tokens from 2.76 million Nepali webpages) as used to train any other model, allowing me to produce an embedding more comparable in scale to English-language GloVe. Training hyperparameters adhered closely to best practices for GloVe.

I also study the following pretrained **generative language models**:

- **OpenAI GPT2-XL**, an English-language model trained on OpenAI's WebText dataset [324].
- **Meta LLaMA-2-7B**, an English-language model trained on public datasets including The Pile [141].
- **DistilGPT2 Nepali**, an open-weight, reduced-parameter version of GPT2 pretrained on the nepalitext dataset, which consists of Nepali text from the CC100 [428] and OSCAR [377] datasets, as well as Nepali Wikipedia.

I use 4-bit quantization [107] to mount LLaMA-2-7B on affordable GPU hardware.

3.4 Methods

I use mixed quantitative and qualitative methods to collect and analyze the presentations of adolescence in AI and those reported by adolescent participants in the study.

3.4.1 Computational Methods

I obtained data from the static word embeddings and generative language models by employing methods appropriate to the models' pretraining objectives.

Static Word Embeddings

For each static word embedding, I computed 1) the 1,000 words *most* associated with adolescents; and 2) the 1,000 most frequently occurring words *uniquely* associated with adolescents over any other age group. Given an embedding vocabulary V, I define an Adolescent target group A.

A, Teenager: *teenager, teenagers, teen, teens, teenage, teenaged, adolescent, adolescence*
To obtain the *most* associated words with A, I compute the mean cosine similarity $s = \frac{\sum_{a \in A} cos(\vec{w}, \vec{a})}{|A|}$ for every word vector $\vec{w}$ corresponding to a word $w \in V$, and select the words with the 1,000 largest values of s.

To obtain the highest frequency words *uniquely* associated with A, I use a Single-Category Word Embedding Association Test (SC-WEAT) [68, 67] to compare the relative similarity of a word w to two attribute groups A and B:

$$d(w, A, B) = \frac{\text{mean}_{a \in A} \cos(\vec{w}, \vec{a}) - \text{mean}_{b \in B} \cos(\vec{w}, \vec{b})}{\text{std\_dev}_{x \in A \cup B} \cos(\vec{w}, \vec{x})} \tag{3.1}$$

The SC-WEAT returns an effect size (Cohen's d) and a p-value based on a permutation test. Unlike some SC-WEATs, which define A and B based on two poles of a binary (*e.g.*, Male/Female), Teenager has no clear opposing pole for B. Thus, I define three B groups using the age ranges specified by the NIH: Children (B_1), Adults (B_2), and Older Adults (B_3):

- B_1, Children: *child, children, childlike, childhood, kid, kids, schoolchild, schoolchildren*
- B_2, Adult: *adult, adults, adulthood, middle-age, middle-aged, grownup, grown-up, grownups*
- B_3, Older Adults: *aged, aging, older, old-age, elder, elders, elderly, retiree*

For every word $w \in V$, an SC-WEAT is taken between A and B_1; A and B_2; and A and B_3. I select only words that exhibit large, statistically significant effect sizes with A when

compared with all three of B_1, B_2, and B_3. Formally, let W_1 denote the words w in V such that $d(w, A, B_1) > 0.8, p < .05$; W_2 the words such that $d(w, A, B_2) > 0.8, p < .05$; and W_3 the words such that, $d(w, A, B_3) > 0.8, p < .05$. That is:

$$\begin{aligned} W_1 &= \{w \in V \mid d(w, A, B_1) > 0.8, p < .05\} \\ W_2 &= \{w \in V \mid d(w, A, B_2) > 0.8, p < .05\} \\ W_3 &= \{w \in V \mid d(w, A, B_3) > 0.8, p < .05\} \end{aligned} \tag{3.2}$$

I select the words W_A exhibiting a large, significant effect with Adolescent over Children, Adults, and Older Adults:

$$W_A = W_1 \cap W_2 \cap W_3 \tag{3.3}$$

I then select from W_A the 1,000 most frequently occurring words in the corpus that produced V, a straightforward task because static word embeddings are rank-ordered based on word frequency. Word groups were constructed by 1) referring NIH descriptions of each age range; and 2) using WordNet [246] to increase the number of words in each group to eight, meeting the SC-WEAT minimum [67]. The Nepali-speaking first co-author translated A, B_1, B_2, and B_3 into Nepali for use with the Nepali embeddings.

Generative Language Models

I study generative language models by using them to generate text conditioned on a prompt. Table 3.1 includes the prompts I designed, drawing on the prior work of Stern (2005) [372], who examined media portrayals of the behaviors, motivations, and relationships of adolescents. Prompts are designed to be 1) consistent with the model's pretraining objective; 2) non-leading and possible to answer in an unbiased manner; and 3) easily adaptable for the human subjects study described below.

To generate text, I use multinomial sampling with the temperature set to 1.0, allowing the model to sample next words based on its probability distribution over the output vocabulary

Domain	Generative Language Model Prompts
Behavior	At school, the teenager
	At home, the teenager
	At work, the teenager
	At the party, the teenager
	Online, the teenager
Motivation	The teenager worked because they wanted
	The teenager studied because they wanted
	The teenager flirted because they wanted
	The teenager socialized because they wanted
	The teenager volunteered because they wanted
Relationship	With their friends, the teenager
	With their parents, the teenager
	With their teachers, the teenager
	With their coworkers, the teenager
	With their romantic partner, the teenager

Table 3.1: Prompts for generative language models, drawing on Stern (2005) [372].

[184]. This allows me to generate 15 distinct continuations for each prompt (225 per model) that are high-probability for the model and representative of its semantic associations. Generative language models are restricted to produce no more than 50 new tokens (words or subwords) of output.

3.4.2 Workshop Sessions

I held workshops on Zoom with N=14 English-speaking adolescents in the U.S. and N=18 Nepali-speaking adolescents in Nepal. My university's IRB approved this study.

Participants

I used purposive sampling [69] to recruit two populations of participants: English-speaking adolescents between 13 and 17 residing in the United States, and Nepali-speaking adolescents between 13 and 17 residing in Nepal. To recruit U.S. participants, my first co-author and I used a contact list of parents who indicated their willingness to be contacted by my university regarding enrolling their children in research. My first co-author and I sent one email to individuals whose children met the study's inclusion criteria, then called them once at the phone number provided. To recruit Nepalese participants, a relative of my co-author residing in Kathmandu posted recruiting flyers at two Kathmandu high schools. The first co-author collected signed assent forms from participants and signed consent forms from their parents. U.S. participants received $25 Amazon credit. Because Amazon does not operate in Nepal (nor does any equivalent), I compensated participants in Nepal via direct payment equal to $7.50 USD in Nepalese Rupees, after consulting a relative of the first co-author living in Nepal regarding exchange rate to ensure I did not bias participant responses [248].

Workshop

All workshops took place over Zoom during December 2023 and January 2024. Participants could choose a synchronous or asynchronous format. With exception of a session wherein two participants asked to join a workshop together, my first co-author and I conducted workshops individually to allow participants more opportunities to ask questions. Sessions began with a five minute, story-based introduction to how AI learns language—for example, by guessing the next word in a sentence, or arranging words based on their similarity to each other. Participants were then asked to help AI learn about teenagers, which involved the following tasks:

- Write the top ten words that come into your head when you hear the word *teenager*.
- Write ten words that *only* describe teenagers, and do not describe children, adults, or

older adults.

- Complete the sentence with a few words, using the generative language model prompts provided in Table 3.1.

- Rate 20 traits on a scale from 1 (most similar) to 5 (least similar) based on how well they describe teenagers.

- Provide the AI with instructions on how to discuss teenagers fairly (both accurately and without bias).

Participants were asked to write about whether and why AI should learn about teenagers from teenagers themselves, rather than media sources. Finally, my first co-author and I engaged in dialogue with synchronous participants to answer their questions about AI. Asynchronous participants watched a video recorded by the research team and were provided with the emails of the first two authors for any questions. U.S. participants completed the research instruments using a Google Form, while Nepalese participants used paper and sent photos to the authors, who transcribed them for further analysis.

3.4.3 Data Analysis

I followed a Directed Content Analysis methodology [15] to analyze data from models and participants. I first used k-means clustering on the word vectors most associated and uniquely associated with adolescents in the GloVe-CC, GloVe-NE, FastText-CC, and FastText-NE embeddings. I selected the number of clusters (between 5 and 10) using Silhouette Score [349]. My first co-author and I then individually reviewed the clusters and assigned labels (*e.g.*, a cluster containing *Justin, Morgan, etc.*, was assigned *First Names*). We then met to discuss and formalize labels into initial codes. We then applied the codes to the generative language model outputs. Where an output did not belong to any existing code, it was added to an *Other* category. After coding the output of each model, we met to review outputs classified

as *Other*, and decided whether to add new codes. We discussed output on which they did not agree and either resolved the code in discussion or added it to the *Other* category if agreement was not reached.

Next, we applied the codes to participant workshop data, adding codes as needed and keeping track via memos of how participant responses differed from model outputs. We sequentially reviewed the word similarity, prompt continuation, and instructions for AI fairness data, meeting to discuss and resolve differences after each phase of coding. All data was coded in Google Sheets, and each author was provided with separate copies of model and participant data so that we could not see each other's codes before discussion. My co-author translated Nepali content and provided guidance where the meaning of a translation was uncertain. After arriving at a final hierarchy of 40 codes with 10 top-level codes such as *Teen Experiences* and *Law and Crime*, we reviewed all the data again, refining code assignments as appropriate.

We then met three times to arrive at themes describing the findings. During the first meeting, we used affinity diagramming to visualize proposed themes that were shared across languages and data sources (model or human) and those which were distinct across languages and sources. After this meeting, we wrote memos describing the proposed themes. We shared the memos and discussed them in the second meeting to arrive at the final themes reflected in the Results.

3.5 Results

Results show biases in static word embeddings and generative language models reflective of the traditional media sources on which they trained, and data from workshops shows AI is misaligned with adolescent life, and adolescents are themselves aware of media biases.

3.5.1 Static Word Embeddings

Table 3.2 illustrates teenage life in clusters of words most-associated and uniquely associated with adolescents. Some clusters are descriptive, with words that mean *teenager*, words related

Most Associated Words (English)				***Most*** Associated Words (Nepali)		
E	%	**Cluster Name**	**Representative Words**	%	**Cluster Name**	**Representative Words**
FT	14.7	Teenagers	teenagers, youths, juveniles	9.0	Teens (female)	young woman, young girl, woman
	12.4	Teen Years	19-year-old, fifteen-years-old	8.2	Teens (male)	adolescent, youthful, young man
	9.5	Other Ages	college-student, baby-boomers	1.1	Age Groups	adult, child, elderly, very young
	8.1	School	high-schooler, middle-schooler	15.0	Teen Names	Surkishore, Ranjeeta, Amritraj
	6.6	Puberty	puberty, pimples, gawkiness	34.2	Life Changes	puberty, menstruation, employable
	10.0	Coming of Age	coming-of-age, prom-night	27.6	Relationships	lovers, friends, mother-son
	8.8	Stereotypes	acne-ridden, braces-wearing	4.8	Cultural Figures	princess, divine girl, Sukanya
	9.7	Rebellion	rebellious, angst-filled			
	1.6	Delinquency	delinquents, runaways, juvey			
	18.6	Sex	barely-legal, underage, jail-bait			
GloVe	11.2	Age Words	16-year-old, youngster, prodigy	17.5	Teenagers	young women, young girl, junior
	8.5	Relationships	dad, mom, friends, lover, teacher	15.3	Relationships	father, son, couple, brother
	15.6	Stereotypes	jocks, nerd, emo, punks, stoned	4.9	School	school, class, principal, studious
	12.0	Mental Illness	self-esteem, psychotic, suicidal	21.7	Names	Rana, Lalit, Mohan, Uttam
	11.8	Risks	at-risk, dropout, pregnancies	11.8	Times	morning, year, Magh (month)
	18.6	Violence	violent, bullied, victim, murder	10.1	Violence	fugitive, murder, kidnapped
	13.1	Sex	horny, masturbating, kinky	18.6	Public Events	demonstration, committee
	9.2	Sexual Taboo	taboo, underage, lolita, voyeur			
Exclusively Associated Words (English)				***Exclusively*** Associated Words (Nepali)		
E	%	**Cluster Name**	**Representative Words**	%	**Cluster Name**	**Representative Words**
FT	16.0	First Names	Sam, Justin, Morgan, Madison	9.8	Internet	URL, Photos, Yahoo, interface
	21.5	Places/Headlines	Seattle, Campus, Driver, Youth	58.3	Travel/Tourism	attractions, architecture, Janakpur
	5.8	Teen Media	vampire, manga, YA, zombies	25.6	Media/Names	BBC, Youtube, Times, Pramod
	5.5	Technology	webcam, Facebook, Instagram	4.0	Technology	Google, Maps, button, lite, free
	27.7	Violence	violent, killer, arrest, shooting	2.3	Years	1977, 1972, 1965, 1963, 1923
	18.2	Drugs/Rebellion	drugs, alcohol, rebel, band, DUI			
	5.3	Sex	sex, porn, breasts, lust, panties			
GloVe	18.1	Sex	sex, erotic, orgasm, porn, incest	21.3	Infrastructure	infotech, grid, construction, metro
	13.9	Sex (Headlines)	Sexy, Naked, BDSM, Lesbian	44.8	Politics	Dharmashala, anti-government
	9.0	Violence	violent, suspects, felony, rape	.01	Music	mixing, mastering
	29.8	Technology	cellphone, clicks, streaming	1.0	Entertainment	Pathao, Tootle, Cartoonz, heroes
	29.2	Celebrities	Rihanna, Spears, Olson, MTV	32.7	Sports Names	Baniya, Neupane, Ashutosh

Table 3.2: Clusters of the most and exclusively associated words with the Teenager group in English and Nepali embeddings.

to school, common names of teenagers, and words for adjacent concepts like other age groups (*baby-boomers*). I derived four themes from static word embeddings.

Instability and Stereotypes

Among the most associated words in English static word embeddings, there exist clusters of stereotypical descriptions (*acne-ridden, braces-wearing, spiky-haired*), media stereotypes (*jocks, nerd, emo*), and words connoting mental illness (*self-esteem, psychotic, suicidal*). A teenage rebellion cluster further illustrates the extent to which adolescents are seen as not in control of their desires, with words such as *sex-crazed* and *drug-crazed*. A similar Drugs & Rebellion cluster forms among the uniquely associated FastText words, highlighting teen drug and alcohol use. These associations find little analogue in Nepali static word embeddings, which lack comparable associations with stereotypes and instability.

Violence and Vulnerability

Risk and violence emerge in the English static word embeddings. Words like *victim* and *at-risk* indicate teenage vulnerability to violence, while *killer* and *suspects* suggest teenagers as perpetrators. Violence takes forms from bullying, to lethal violence such as *murder* and *suicide*, to sexual violence including *rape*, to criminal violence (*arrest*, *felony*), to sensationalized violence like *torture*. Violence composed the single largest cluster of uniquely associated words (27.6%) in the English Fasttext embedding. I identified a Violence cluster in the most associated Nepali GloVe words (*fugitive, murder, police*), but it is smaller than English Violence clusters and mostly free of sensationalized violence.

Sex and Sexualization

Sexual taboo and fetishization of adolescents emerge in the most and uniquely associated words in English static word embeddings. Words like *lolita, underage, barely-legal*, and *jail-bait* occur in the most-associated words, along with *voyeur*. The word *porn* occurs among uniquely

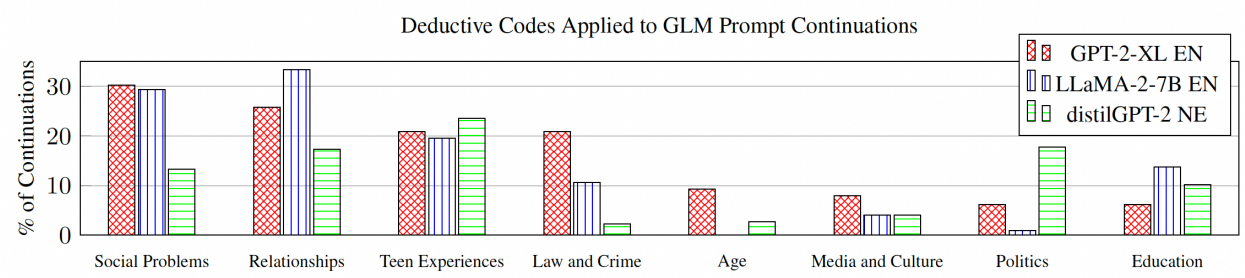


Figure 3.2: Social problems predominate in the English-language generative models continuing prompts related to teenagers.

associated words, along with a cluster of capitalized words including (*BDSM, Lesbian, Naked*), suggesting an origin in the headlines of pornographic webpages. Pornographic and fetishizing clusters are distinct from clusters of sexual desire words, which occur in Nepali and English static word embeddings and include words like *lust, sexual pleasure,* and *lovers*.

Emerging Adulthood

The English FastText embedding includes a Coming-of-Age cluster (*coming-of-age, right-of-passage*), while clusters related to the bodily transition of puberty occur in English static word embeddings (*puberty, gawkiness*) and Nepali static word embeddings (*puberty, menstruation*). The Nepali FastText cluster also includes words related to adult roles in marriage and work (*marriageable, employable*). Though it was not apparent until interacting with Nepalese adolescents, Infrastructure (*infotech, construction*) and Public Events (*demonstration, program*) clusters also point to emerging adulthood, as adolescents can graduate from high school after 10th grade and take a job in a trade, beginning adult life.

3.5.2 Generative Language Models

Figure 3.2 visualizes the deductive codes applied to the 225 continuations from three generative language models, based on the prompts in Table 3.1. I derived three themes for generative

language model outputs.

Social Problems—Especially Violence—Are Common

30% of GPT-2-XL continuations and 29% of LLaMA-2 continuations received the Social Problems code, making this the most common code for GPT-2-XL and the second most common for LLaMA-2. Of the Social Problems continuations, 47% were subcoded for Violence in LLaMA-2, and 50% in GPT-2-XL. For example, the following was generated by LLaMA-2 from "At home, the teenager": *was bullied by his mother's boyfriend. At school, he was taunted by the kids. He was so depressed, he attempted suicide.* Other common subcodes included Drug Use (21% LLaMA-2, 9% GPT2-XL); Teen Trauma (17% LLaMA-2, 21% GPT2-XL); Mental Illness (9% LLaMA-2, 12% GPT2-XL); and Sexualization (9% LLaMA-2, 13% GPT2-XL), as in this continuation from GPT2-XL: "Online, the teenager": *was charged with child porn and illegal computer access. After the investigation was closed into his alleged illegal access, a case had to be filed.* Though much less common, violence also occurs in the continuations of DistilGPT2-Nepali. Bullying is absent, but suicide and sexual violence occur in the roughly 2% of continuations coded as Law and Crime. Though social problems are the default in English, one also observes teenage exemplars—noteworthy exceptions to the norm. For example, LLaMA-2 continues "At school, the teenager" with *has a very good academic record, and is a member of the student council. In addition to her school duties, she has been a member of the Girl Scouts since she was in the first grade.*

Sensationalism Emerges from "High-Quality" Training Data

Many generative language model continuations, including those resulting in social problems and violence, either 1) followed a distinct journalistic style or 2) explicitly cited a news media source or described a quote being taken by a media source. The following representative example from LLaMA-2-7B was generated from "At school, the teenager": *was bullied for his sexual orientation. The 15-year-old boy from the village of Nizhny Novgorod, who was bullied for his sexual orientation, committed suicide.* The continuation follows a journalistic

style that concisely communicates the boy's age, hometown, and circumstances leading to the events under consideration. In other cases, the model appears to shift into a journalistic mode of writing; LLaMA-2 continues "The teenager flirted because they wanted: *to have sex with her. A 17-year-old girl from Warrington has been found guilty of having sex with a 14-year-old boy.* Other continuations identify quotes taken by media outlets, including CNNMoney, KRIV-TV, and the Daily News. In one case, a LLaMA-2 output noted that photos were provided by Getty Images. Continuations by DistilGPT2-Nepali often included the apparent source of the model's continuation, such as Everest Online News, eHimala, and Federation of Nepal Journalists. Even models trained on reputable sources of text data are vulnerable to sensationalism and societal bias, if reflected in the media.

Societally Sanctioned Activities for Adolescents

The codes appropriate to generative language model continuations also surfaced societal attitudes toward specific adolescent activities. Prompts about parties were the most likely to result in continuations involving social problems, followed by prompts about teenagers online. Prompts about teenagers in the workplace were *least* likely to produce continuations involving societal problems, although many English-language continuations trivialize adolescent work; for example, several LLaMA-2 continuations discussed adolescents being fired for refusing to take drug tests. Prompts about school were the most likely to be coded for adolescent relationships, while prompts involving the home were the most likely to involve adolescent experiences, as in the LLaMA-2-7B continuation of "At home, the teenager": *is a person who is looking for their identity. They are trying to find out what they are about.*

3.5.3 Workshop Sessions

Workshop data demonstrates that AI reflections of teenage life are disconnected from the experiences of adolescents. I derived three themes from participant responses.

Most Similar Words (U.S. Participants)			***Most*** Similar Words (Nepalese Participants)		
%	**Cluster Name**	**Representative Words**	%	**Cluster Name**	**Representative Words**
10.7	Fun	fun, party, fashion, curiosity	23.5	Energy	energetic, playful, excited, emotional
12.0	Stress	stress, moody, rebellious, reactive	26.5	Stress	stress, pressure, fear, gossip, angry
12.0	Immaturity	immature, irresponsible, insecure	10.3	Immaturity	immaturity, shy, ignorant, fake
20.0	Discovery	discovery, growth, independence	7.4	Innocence	childhood, innocent, obedient, sleepy
20.0	Social Life	social, friendly, family, bonds	32.4	Likability	friendly, cool, beautiful, youth
12.0	School	grades, homework, procrastination			
8.0	Boredom	bored, lazy, dull, tired			
5.3	Difference	different, makeup, sleep, phone			
Exclusively Similar Words (U.S. Participants)			***Exclusively*** Similar Words (Nepalese Participants)		
%	**Cluster Name**	**Representative Words**	%	**Cluster Name**	**Representative Words**
18.9	Uncertainty	questioning, overthinking, impulsive	15.3	Pressure	pressure, showoff, drama, ruthless
26.4	Change	changing, different, curious, frisky	20.8	Freedom	freedom, independent, creative
15.1	Impatience	impatient, restless, reckless, moody	19.4	Impatience	restless, irritation, unsatisfied, greedy
22.6	Inexperience	confused, misunderstood, inexperienced	8.3	Inexperience	uninformed, shy, lazy, solitary
17.0	Eagerness	idealistic, impressionable, attentive	9.7	Adventure	adventurous, excited, expressive
			16.7	Likability	chill, clever, fashionable, good
			20.8	Discipline	disciplined, work, study, attitude

Table 3.3: Clusters of words associated with teenagers, according to teen participants in the U.S. and Nepal.

AI Does Not Reflect Adolescent Views of Adolescence

As discussed in the Methods, participants rated 20 trait words (*e.g.*, *opinionated, thoughtful*) from 1 to 5 based on how well they described teenagers. I took the same words and computed the cosine similarity between the $\vec{teenager}$ vector and the trait word vector. I then took the correlation between mean participant ratings and cosine similarities, obtaining Pearson's ρ=.02, *n.s.* for English FastText, and ρ=.06, *n.s.* in English GloVe, indicating no correlation between static word embeddings and human ratings, as shown in Fig 3.3. Similar results were obtained for Nepali embeddings, with ρ=.06, *n.s.* in Nepali FastText, and ρ=−.23, *n.s.* in Nepali GloVe.

As shown in Table 3.3, I also clustered the most-associated and uniquely-associated words provided by teenagers, using a vector for each word based on its valence, arousal, and dominance in the lexicon of Mohammad (2018) [251], and applying the k-means algorithm. U.S. clusters suggest a strikingly different view of adolescent life than that of English static

word embeddings. Clusters related to School, Social Life, Discovery, and Fun make up more than 60% of the clustered most similar words. Where more negative traits like *rebellious* and *insecure* emerge, they are balanced by apparent explanations suggested by words like *stress* and *anxiety*. Clusters of exclusively associated words bear more resemblance to English static word embeddings, with Change and Uncertainty making up more than 45% of the clustered words. However, the clusters also surface feelings of Inexperience (*confused, misunderstood, gullible*) and Eagerness for the future (*idealistic, attentive, college*). Notably absent is *any* word connoting violence or lurid sexuality. Nepalese exclusively associated clusters similarly describe Impatience, Inexperience, and interest in Freedom and Adventure. Clusters related to Likability (*cool, beautiful, chill, clever*) occur in both the most and exclusively associated words, while words related to Pressure and Discipline, with a particular focus on school (*disciplined, study, pressure*), make up more than 35% of the clustered exclusively associated words.

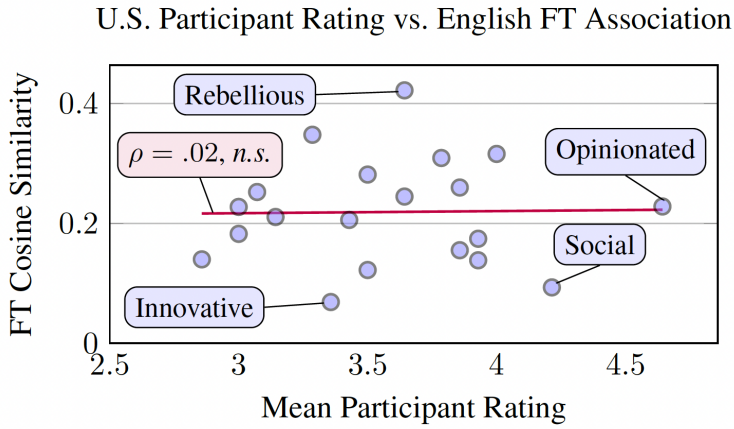


Figure 3.3: Word associations with "teenager" in FT are decorrelated from U.S. teens' ratings of similarity to "teenager."

Adolescent Life is Not Well-Characterized by Newsworthy Events

Qualitative analysis showed that participant prompt continuations were misaligned with the continuations of generative language models. Prompted with "At school, the teenager", U.S. participants responded with *writes in a notebook* (E8), *doesn't pay attention to the teacher* (E1), *studies in class* (E12), and *eats lunch* (E5). Prompted with "At home, the teenager", four U.S. participants wrote about videogames, three about sleeping, and two about homework. Videogames and watching online videos on platforms like Tiktok also constituted the majority of responses to the prompt "Online, the teenager." Six continuations of "At the party, the teenager" included talking to friends, while two discussed drinking alcohol. Aside from one mention each of cyberbullying and shoplifting, participant continuations are devoid of violence, rebellion, and sexualization. A far cry from the social problems in generative language models, the only description of a teenager facing discipline is specified by E10 for "With their teachers, the teenager": *got in trouble for sleeping in class.*

Responses from Nepalese participants were similarly mundane. Continuing "At school, the teenager," nine participants described studying, learning, or reading, two described respecting teachers, and two described getting scoldings or beatings from teachers. In response to "At home, the teenager", five participants described doing chores, three using a cellphone, two browsing social media, and three doing homework. In response to "At the party, the teenager," five participants described dancing, three wearing new or beautiful clothing, and three eating or feasting. In response to "Online, the teenager", six participants described searching for information or studying, five chatting or gossiping, and two playing games. Far from the sensationlized outputs of static word embeddings and generative language models, adolescents describe everyday activities: going to school, playing videogames, and talking with friends.

Societal Expectations Inform Adolescent Presentations of Adulthood

Comparing responses of U.S. and Nepalese participants revealed differing manifestations of emerging adulthood. Responding to "At work, the teenager", eight Nepalese participants

wrote that the teenager is hardworking, while three others described focusing, or being fired due to lack of focus. In response to "The teenager worked because they wanted", seven participants described a *shortage* or need of money, and two more described helping with family finances. By contrast, every U.S. participant wrote *money*, describing potential uses of this money to buy clothes (E9), new games (E10), a car (E11), or just *stuff* (E1, E12). E3 wrote *the freedom that money allows while having minimal bills.* Responding to "At work, the teenager", three U.S. participants described completing assigned tasks, two talking to friends or coworkers, playing on their phone (E11), ignoring their manager (E13), or doing the *bare minimum* (E1). Where U.S. participants described work as an avenue to independence and agency, Nepalese participants described it as a means of supporting their family. Both descriptions reflect emerging adulthood, contextualized by the expectations and opportunities of two societies.

3.5.4 Instructions for Fair AI

Participants wrote instructions for AI to represent teenagers fairly, and shared thoughts on the sources of data on which AI trained. I arrived at four themes based on this data.

Adolescents are Aware of Media Stereotypes

U.S. participants contended that media representations of teenagers are biased and reflect a stigma around adolescence. E7 wrote: *Out of all age groups, teenagers are by far the most stigmatized* and *many people hold stereotypical views of teenagers... consistently reinforced through media.* Similarly, E4 wrote *teenagers are viewed in a very negative light because we have a tendency to deal with things in a very different way than adults or people from other age groups deal with their problems.* Nepalese participants also highlighted that societal views differ from those of teenagers. N16 wrote that it is *important to describe the teenager as they are... teenagers' views are different from society's point of view.* N13 wrote *teenager[s] aren't like the society think[s,] because they create their own way.* Participants also noted that *how* AI learned about adolescents would affect their view of using it. E8 wrote: *for teenagers*

to feel seen or heard I think it would be good to have them be the ones that tell [AI] about themselves and not have [it] assuming. E6 wrote that, were AI to train on *data on teenagers from the media, [it] would most likely learn what a stereotypical teenager is like and not how they actually are. The media usually puts teenagers in a bad light but... they can be smart, well mannered, and successful.* E10 wrote that AI trained on media would be disconnected from teenage life, noting *Teens make fun of how movies and TV shows portray them, finding it to be really far off from what they are in real life.* Finally, N13 wrote *AI should represent [teenagers] as they are rather than what other[s] think of them.*

No Media Source is Unbiased, But Some are More Biased Than Others

Reflecting on using traditional and online media sources for AI training data, E11 wrote: *movies, newspapers, and other media often portray teens in a stereotypical fashion that only captures part of what a teen really is. The information... would be surface level at best.* E13 wrote that if AI systems *read the newspaper, much of the information they would gain could be false as it is the way others view teenagers rather than the way they actually are. Whereas teenagers would be able to provide the real way they see themselves.* N4 stressed the disconnect between media and reality, writing *what we learn from media and newspapers is different [from] when we learn from human beings.*

Participants acknowledged that perfectly unbiased media might be unachievable. E1 wrote: *I think it is almost impossible to represent teenagers, or anything really, in media without some kind of bias.* E9 further noted: *the way social media represents teenagers can be very far-fetched, and possibly even offensive to what teenagers are really like. I believe it's important for... AI to accurately represent teenager[s] in comparison to possible lies and fake information being spread about them. But... all teenagers are different so I don't believe there's a specific way to represent them all accurately.* E3 highlighted that the attention-driven business model of media companies underlies the problem, writing *I don't think the media is a good representation of any group of people because of the business model they work under.*

Most participants agreed that AI should interact with teenagers to learn about them. N17

wrote: *Teens know more about themself than [any] other. So if teenagers teach [AI] about them it will be more effective compare[d] to learning about them from other media.* N1 wrote: *media only explains about surface feeling[s,] but a teenager could explain about it in detail.* Finally, E10 suggested that AI might *search through past chats with other teens in order to figure out what shared interests most teenagers have,* a strategy similar to that employed by many chat-based language models, which train on datasets of conversations [480]. While such a dataset might raise an array of ethical concerns, E10 identifies a gap in training data for conversational models specific to underrepresented user groups.

Diversity and Positivity: Perspectives on Fair Representation

Two perspectives on how adolescents could be fairly represented by AI emerged in the data. U.S. participants (nine of thirteen) stressed portraying the *diversity* of teenagers. E7 wrote: *Instruction 1: Clarify that not all teenagers are the same. As it is with every age group, traits can vary drastically between individuals.* E3 wanted to ensure that AI would *include examples of teenagers from all backgrounds.* E9 noted: *teenagers are all very different. . . there's no specific category to place teenagers under.* The preference for diverse representation was sometimes juxtaposed with an assumption that AI would focus on adolescents' negative traits. E1 wrote: *Instruction 1: When asked about teenagers, don't just say the bad things; teenagers are different from each other, so you should represent all of them.* E13 wrote: *Give both good and bad examples. For example, mention that they are rebellious but also innovative.* Where U.S. participants stressed diversity, Nepalese participants centered *positivity*, with ten participants listing positive traits in instructions to AI. N9 wrote that AI should reflect that *teenagers are the most creative and confiden[t] and thoughtful.* N13 similarly wrote that teenagers *are free minded, introvert[ed], and curious.* While the preference for diversity may reflect a U.S. cultural value, the motivation is similar between U.S. and Nepalese participants: to present adolescents generously, including positive traits rather than replicating negative media biases.

The Potential for AI to Correct Stereotypes

Both U.S. and Nepalese participants expressed optimism that AI could help in correcting stereotypes. E10 positioned AI as a mediator, writing that *society has a negative stereotype of teenagers, that they are moody for no reason and that they are disrespectful. But teens have various reasons for acting the way they do, and [AI] could help people understand that.* E13 suggested proactively addressing biases, writing *there is no way to break the social stereotype that teenagers act a certain way if the only information being put out about teens supports the stereotype, rather than showing the stereotype is false.* N4 wrote that *AI could express the teenagers in [a] way [that] every one will accept it.* Highlighting that AI could serve as a vector for better interpersonal communication, N7 said that *society should also know about how the teenagers feel and the way they think.* In contrast with existing information architectures like social media, N1 wrote that *AI could be the place where teenagers feels safe.*

3.6 Discussion

I show that even training on high-quality data sources like news articles can reproduce harmful societal attitudes depicting adolescents as violent, criminal, and rebellious. That some of these biases do not exist in monolingual Nepali-language models might prompt us to re-examine assumptions that these biases are unavoidable. Moreover, that user-facing generative language models associate adolescence with social problems shows the potential for AI to amplify bias, as it serves as a mediator of culture [56, 99] and a source of information [242].

Adolescents' access to information and shared spaces is often mediated by societal attitudes. For example, Bernier (2011) [31] find that only 2.2% of facility square footage is devoted to teenage users in libraries, where youth represent nearly 25% of all users, observing that this disparity is motivated by unsavory stereotypes and marginalizes them in a space for information seeking. As AI begins to serve society's information seeking needs, this work poses the question of whether AI can serve as a *place where teenagers feel safe*, as N1 put it,

or if it will reflect the attitudes and serve primarily the needs of adults. Feeling safe using AI may also support teen development by providing a space to "enact maturity," inviting teens into conversations about consequential subjects, like politics [21].

Participants also saw AI as a means of addressing societal stigma in traditional and social media. To do so, they believed AI would need to understand adolescents by interacting directly with them. Some participants even envisioned AI mediating between adolescents and adults, providing perspective when teens aren't able to express themselves. Such optimism about the role of AI suggests the need to develop frameworks for ethical engagement between adolescents and language technologies. While AI may hold potential for changing societal attitudes, it can also be used to collect data or financial resources from users [441]. Finding ways to maximize user agency while personalizing models could be explored in future work.

The study paired an analysis of a societal attitude in AI with a human subjects study of the group impacted, revealing the disconnect between adolescent experiences of the world and AI presentations. Participants provided context that helped to understand how societal expectations of teenagers shape their self-presentation, and their presentation in media sources. This work indicates that more complete descriptions of AI and societal biases can be obtained through mixed methods work, involving not only AI-based measurements but also participation of human subjects.

3.6.1 Limitations and Future Work

I used solely monolingual, open models to maximize reproducibility and prevent cross-lingual transfer of semantic associations. Nonetheless, most users prefer proprietary, chat-based, multilingual models like ChatGPT. Future work might examine such models not as reflections of culture but as sociotechnical tools. Moreover, while the Nepali-language models used are the best I know of, I observed some disfluencies in their output, a limitation for low-resource languages. Finally, adolescents are a large, diverse group that I cannot hope to fully capture in a single study.

3.7 Conclusion

This work demonstrates the epistemic risk introduced by relying on media sources to inform how groups of people are represented in generative and general-purpose AI. It showed that the lurid and sensationalized depictions of adolescents present in AI are decoupled from the everyday experiences of U.S. and Nepalese adolescents, whom the workshops revealed are well-aware of media stereotypes. However, it also offers a means of mitigating that risk by centering the perspectives of those affected: even as teenagers grapple with perceived social stigma, they view AI as having potential to help create a safer and more positive environment for adolescents. I hope this research will inspire further work that helps to realize that goal.

Chapter 4

SCALING PERFORMANCE, SCALING RISK: DATASET SCALE AND SOCIETAL CONSISTENCY MEDIATE FACIAL IMPRESSION BIAS IN LANGUAGE-VISION AI

4.1 Preface

One of the central contentions of AI research over the past ten years has been that many problems are solvable via *scaling*: all else being equal, increasing the parameter count of a model or the size of the dataset on which a model trains will result in corresponding improvements in the model's performance [58, 84, 171]. In some cases, increasing scale appears to facilitate the emergence of novel model capabilities, such as few-shot learning in GPT models [324, 58]. But how can researchers and practitioners understand the epistemic risk induced by datasets that are too large for curators to audit or meaningfully describe [28], and by the ever more complex models fit to those datasets? In this study, I consider the role played by scale in amplifying the epistemic risk posed by general-purpose vision-language AI. The findings of this chapter support **Thesis Statement 2**: *Epistemic risk in general-purpose models is predicted by 1) the size of the dataset on which a model has been pretrained and 2) the relative consistency of the bias, insofar as this can be captured by large-scale psychometric surveys.*

4.2 Introduction

OpenAI's multimodal GPT-4 powers the beta version of Be My AI, an extension of the Be My Eyes app [24, 25] that provides "instantaneous identification, interpretation, and conversational visual assistance" to blind and low-vision users. Until recently, the app allowed users to ask questions about images of people and receive live explanations. The temporary

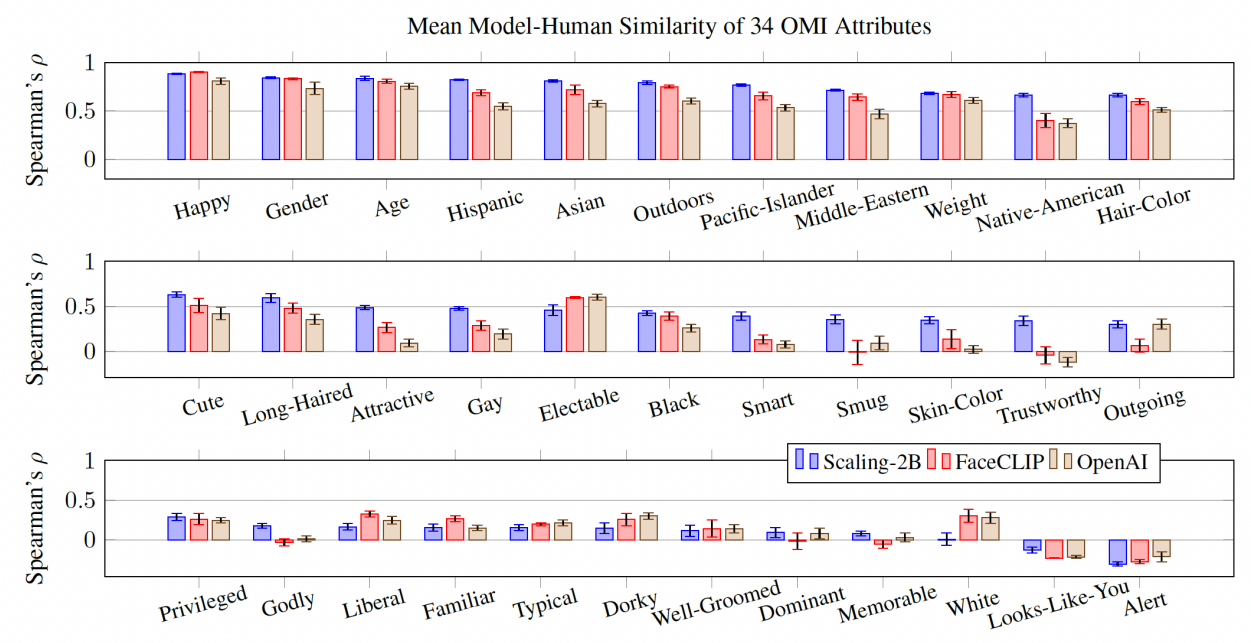


Figure 4.1: CLIP models learn human-like facial impression biases. The highest model-human correlations are obtained for intuitively visual categories that are broadly shared by a society (such as gender, age, and happiness). Models trained on the largest dataset (LAION-2B) exhibit more human-like biases than FaceCLIP or OpenAI models for most attributes.

discontinuation of this feature was motivated by concern that GPT-4 "would say things it shouldn't about people's faces, such as assessing their gender or emotional state" [173].

The decision belies a broader concern: that by learning to associate language and images, multimodal AI may make insufficiently informed judgments about human attributes based solely on a person's face. When studied in human subjects, this kind of inference is known as a "first impression" or "facial impression" bias [396], and it is known to affect consequential spheres of human social life such as criminal sentencing [431], employment decisions [374], and political elections [11]. Such impressions can include traits like trustworthiness, which are unobservable from a person's face and societally mediated to extent that they are consistent in a population [396]. While psychologists have used computational geometry and supervised

machine learning approaches to modeling facial impression biases [48], it is not known whether semi-supervised vision-language AI models could inadvertently learn such biases in pretraining and propagate them to the many domains in which such models are used.

While features permitting facial image analysis are disabled in GPT-4, the opportunity to study facial impression bias is afforded by CLIP ("Contrastive Language Image Pretraining"), a state-of-the-art vision-language model that allows users to define text classes at inference using natural language [321]. Rather than fine-tuning CLIP to model facial impressions similar to prior work using supervised learning, I study this bias in three families of pretrained CLIP models used in a wide range of multimodal computer vision tasks: the nine models trained by OpenAI [321]; five "FaceCLIP" models post-trained for facial analysis [482]; and 29 "Scaling" models trained by Cherti et al. (2022) [84] on systematically differing amounts of data, allowing for statistical analysis of the effects of model and dataset parameters on facial impression bias.

Analyzing whether CLIP models learn human-like facial impression biases requires a reliable source of human data. This research uses the authoritative One Million Impressions (OMI) dataset of Peterson et al. (2022) [304], which includes 1,004 images of faces rated by human participants across 34 attributes, with which Peterson et al. (2022) [304] learned a supervised model of facial impression biases. In the present work, I used CLIP to compute the similarity of each OMI image to text prompts for the 34 attributes, mimicking the task given to human subjects, and I compared the CLIP similarities to human subject ratings. I offer four primary findings:

1. **CLIP models learn societal facial impression biases, including for unobservable traits such as trustworthiness and sexuality.** Moreover, the extent to which an attribute bias is learned by a CLIP model is strongly correlated with the inter-rater reliability (IRR) of human judgments of the attribute (Spearman's $\rho = .73$ for OpenAI models; $\rho = .76$ for FaceCLIP models; and $\rho = .72$ for Scaling models). A multiple linear regression predicting the similarity of CLIP bias to Human bias finds that the IRR

of the attribute plays a larger role than any model-related variable, with $t(912) = 25.47$, $p < .001$. The extent to which a facial impression bias is learned by a model depends on how consistently it is shared in the population that produced the data on which the models trains.

2. **Dataset Scale is a significant predictor of facial impression bias in CLIP.** Comparison of model-human similarity in two groups of nine CLIP models trained on LAION-80M (80 million examples) and LAION-400M (407 million examples) yields large effect sizes ($d > 0.8$) and statistically significant ($p < .05$) paired samples t-tests for 17 of 34 attributes, indicating increases in the human similarity of bias in models trained on LAION-400M. Differences between models trained on LAION-2B (2.32 billion examples) and LAION-400M are mostly not significant, with the notable exception of *unobservable* attributes like trustworthiness ($d = 1.33$, $p < .05$) and sexuality ($d = 1.14$, $p < .05$). While models trained on larger datasets exhibit stronger task performance [84], they also more faithfully reflect the biases of the population that produced the data.

3. **CLIP models learn human-like associations *between* facial impression biases.** Hierarchical clustering of CLIP and OMI attribute correlation matrices reveals similar groupings of traits, including clusters related to ethnicity, and clusters grouping gender, sexuality, and age. Computing the normalized Frobenius inner product of CLIP correlation matrices with the OMI matrix reveals increasing similarity as pretraining data size increases, with a one-way ANOVA yielding $F(2) = 15.71$, $p < .001$, and large ($d > 0.8$) pairwise effect sizes between groups of Scaling-80M with 400M and 2B models.

4. **Stable Diffusion text-to-image models employing CLIP as a text encoder learn facial impression biases that intersect with demographic biases.** Images generated by Stable Diffusion XL-Turbo (SDXL) are classified by a classifier fit using the OMI images. Classifications reflect human facial impression biases for subjective,

observable attributes like attractiveness (F1=.98), and to a lesser extent for unobservable attributes like liberal (F1=.68) and smart (F1=.65). Applying the classifier to SDXL images generated for White and Black prompts reveals biases in SDXL differentially associating White individuals with traits like memorable, attractive, electable, and happy.

Training VL models on vast web-scraped datasets produces consequential emergent biases. Such models may serve as useful tools for social science, illuminating factors that contribute to human bias. However, the presence of these biases also renders fraught VL models' real-world use, as they may subtly reinforce existing inequities in an online environment increasingly mediated by AI.

4.3 Data

This research uses the One Million Impressions (OMI) dataset and 43 English-language CLIP models trained on web-scraped text-and-image datasets, as well as three text-to-image generators employing CLIP models as text encoders.

4.3.1 The One Million Impressions Dataset

The OMI dataset is a collection of 1,004 images of human faces produced by Peterson et al. (2022) [304] using StyleGAN-2 [200]. Each face is rated by 30 or more human participants on Amazon Mechanical Turk for 34 attributes. For each attribute, participants rate the face on a sliding scale, where one end represents one pole of an attribute binary (such as "trustworthy") and the other end represents the opposing pole of the binary (such as "untrustworthy"). The OMI dataset records the mean participant rating for each of the 34 attributes. Consistent with Peterson et al. (2022) [304], I use these ratings as measurements of human bias at a societal scale, against which CLIP associations can be compared.

Figure 4.2: Examples from the OMI dataset repository at https://github.com/jcpeterson/omi, used as stimuli in this research.

4.3.2 CLIP Training Data

I study CLIP models pretrained on one of five datasets, ordered from smallest to largest:

- **LAION-Face**: A 20-million sample subset of human faces and captions filtered from LAION-400M (see below) using RetinaFace [104] and intended for training facial analysis models [482].
- **LAION-80M**: An 80-million sample subset of LAION-2B (see below) created by Cherti et al. (2022) [84] to study scaling behavior in CLIP.
- **LAION-Aesthetics**: A 120-million sample subset of aesthetically pleasing images from LAION-5B as determined using a CLIP model [358].
- **WebImageText (WIT)**: A web-scraped corpus of 400 million images and captions, constructed by Radford et al. (2021) [321] from a query list using Wikipedia and WordNet.
- **LAION-400M** An open source collection of 407 million image-text pairs intended to replicate the WIT dataset [359].
- **LAION-2B**: An open source English-language dataset of 2.32 billion image-text pairs [358].

4.3.3 Pretrained CLIP Models

This research studies the following CLIP models:

- **OpenAI CLIP**: 9 models pretrained by Radford et al. (2021) [321] on the WIT dataset.

- **Scaling CLIP**: 29 models pretrained by Cherti et al. (2022) [84] on LAION-80M, LAION-400M, and LAION-2B to study CLIP scaling behavior.

- **FaceCLIP**: 5 models trained on the LAION-Face dataset of Zheng et al. (2022) [482]. FaceCLIP models post-train from pretrained OpenAI CLIP-ViT models.

4.3.4 Pretrained Stable Diffusion Models

This research studies three Stable Diffusion (SD) models:

- **Stable Diffusion XL-Turbo**: A high-resolution text-to-image generator employing adversarial distillation diffusion to speed up the rate of image generation [354]. Uses a CLIP-ViT-L and CLIP-ViT-bigG for its text encoder, and pretrains on an internal dataset.

- **Runway Stable Diffusion 1.5**: A high-resolution text-to-image generator finetuned on LAION-Aesthetics [345]. Uses a CLIP-ViT-L-14 as the text encoder, and pretrains on LAION-5B.

- **Stable Diffusion 2**: A text-to-image generator using a CLIP-ViT-H as the text encoder, and pretraining on a filtered subset of LAION-5B [345].

4.4 Approach

I used embeddings from 43 CLIP models to compare facial impression biases measured in CLIP to biases measured in humans by Peterson et al. [304], and extended subspace projection methods from prior work to study bias in generative text-to-image models.

4.4.1 Obtaining Image and Text Embeddings

I obtained image embeddings for the 1,004 images in the OMI dataset after projection to each CLIP model's text-image latent space. Text embeddings use the "a photo of *image class*" prompt recommended by Radford et al. (2021) [321]. Because OMI consists of images of faces, I modify this prompt to "a photo of someone who is *attribute*." In keeping with the binary sliding scale of Peterson et al. (2022) [304], I computed an image's association with each attribute by subtracting its cosine similarity with one pole of the attribute binary ("a photo of someone who has dark hair") from its similarity with opposing pole ("a photo of someone who has light hair"). Formally, given a model m from which embeddings are obtained, the association m_j^a of an image vector $\vec{i_j}$ at index j of the OMI dataset with an attribute a is the difference of the vector's cosine similarity with a positive pole text vector $\vec{t_{a^+}}$ and its cosine similarity with a negative pole text vector $\vec{t_{a^-}}$:

$$m_j^a = cos(\vec{i_j}, \vec{t}_{a^+}) - cos(\vec{i_j}, \vec{t}_{a^-}) \tag{4.1}$$

4.4.2 Adjusting Prompts for Negation

CLIP may fail to adjust for negation in text prompts [296] and can behave like a visual bag-of-words model [461]. For example, CLIP might match the text "a photo with no apples" to a photo of apples, due to how unlikely it is for a text caption (*i.e.,* CLIP's training supervision) to describe something not present in the photo. To adjust for this, negative pole prompts were chosen such that they did not simply negate the positive class. For example, the "outgoing" attribute uses "a photo of someone who is shy" as the negative text class,

rather than "a photo of someone who is not outgoing." This strategy is not viable for some attributes, like those related to ethnicity, which instead use "a photo of someone" as the negative prompt.

4.4.3 Computing CLIP Model-Human Similarity

I denote the ordered set of n=1,004 OMI images as I. The vector of associations $\mathbf{m^a}$ for a model m with an attribute a for all images $i \in I$ is given by:

$$\mathbf{m^a} = (m_0^a, m_1^a, \ldots, m_{n-1}^a, m_n^a) \tag{4.2}$$

Similarly, the vector of human-rated associations $\mathbf{h^a}$ for attribute a for all images $i \in I$ is given by:

$$\mathbf{h^a} = (h_0^a, h_1^a, \ldots, h_{n-1}^a, h_n^a) \tag{4.3}$$

where h_j^a denotes the OMI mean for image $\vec{i}_j$ at index j. The similarity s_m^a of bias in a model m for attribute a to human bias is given by Spearman's ρ:

$$s_m^a = \rho(\mathbf{m^a}, \mathbf{h^a}) \tag{4.4}$$

4.4.4 CAT: Correlated Attribute Test

I compute the correlation between two attributes in a CLIP model using a simple test I call the CAT. As above, the vector of associations $\mathbf{m^a}$ for a model m is given by:

$$\mathbf{m^a} = (m_0^a, m_1^a, \ldots, m_{n-1}^a, m_n^a) \tag{4.5}$$

The measurement $\text{CAT}_m(a, b)$ between attributes a and b in a model m is given by Spearman's ρ:

$$\text{CAT}_m(a, b) = \rho(\mathbf{m^a}, \mathbf{m^b}) \tag{4.6}$$

4.4.5 Subspace Projection for Text-to-Image Models

I draw on subspace projection methods used by Bolukbasi et al. (2016) [50] and Omrani Sabbaghi et al. (2023) [278] to measure facial impression biases in generative text-to-image models. First, I first obtain image embeddings for the 1,004 OMI images from the top layer of a ViT-Large-Patch32-384 model pretrained on ImageNet. For each attribute a, I learn a weights vector w^a predicting the OMI attribute ratings h^a, corresponding to a semantic subspace in the embeddings for the attribute. I then use a generative model g to generate n images via a prompt corresponding to either the positive (a^+) or negative (a^-) pole of an attribute. I embed each generated image g_j at position j with the ViT-Large-Patch32-384 to obtain the vector $\vec{g_j}$, and compute an attribute association g_j^a as its projection product with w^a:

$$g_j^a = \frac{\vec{g_j} \cdot w^a}{||w^a||} \tag{4.7}$$

The vector of associations g^a for a generative model g with an attribute a is given by:

$$\mathbf{g^a} = (g_0^a, g_1^a, \ldots, g_{n-1}^a, g_n^a) \tag{4.8}$$

4.5 Experiments

Four experiments test the existence of human-like facial impression bias in vision-language AI, with consideration given to Human IRR, model and dataset scale, and downstream impact in image generation.

4.5.1 Model vs. Human Biases

I tested whether OpenAI, Scaling, and FaceCLIP CLIP models reflect human-like facial impression biases. I obtained the human-model similarity s_m^a for each model m with each attribute a for the 34 OMI attributes. I then compared the mean human-model similarity for each group of models to the Human IRR for the attribute reported by Peterson et al.

(2022) [304], calculating Pearson's ρ between Human IRR and model-human similarity for each of the 34 attributes. A large coefficient indicates that the more societally consistent a facial impression bias (*i.e.*, as Human IRR increases), the more likely the bias is to be learned during semi-supervised CLIP training. I also computed Pearson's ρ pairwise between OpenAI, Scaling, and FaRL models to assess whether models trained on different datasets learn similar biases.

4.5.2 Effects of Dataset Scale

I calculated human-model similarity s_m^a for each model m with each attribute a for the 34 attributes studied, and I constructed a multiple linear regression to predict the model-human similarity s_m^a for a given CLIP model m and an attribute a. I examined the 27 CLIP models trained by Cherti et al. (2022) [84] and produced via the combination of three CLIP architectures (ViT-B32, ViT-B16, and ViT-L14), three dataset sizes (80M, 400M, 2B), and three total training example counts (3B, 13B, 34B). Independent variables include Human IRR (from Peterson et al. [304]), as well as Dataset Size, Model Parameter Count, and Total Training Examples (from Cherti et al. (2022) [84]). I normalized variables with range outside of $(0, 1)$ by dividing by their max. I conducted post hoc comparisons between each level of scale (80m, 400m, 2b), using paired-samples t-tests.

4.5.3 Structure of Facial Impressions

I computed the correlation matrix $\mathbf{C_m}$ for the 27 CLIP models studied in the Dataset Scale analysis by obtaining $\text{CAT}_m(a, b)$ for every attribute pair $a \in A$ and $b \in A$. I computed a corresponding matrix $\mathbf{C_h}$ by obtaining correlations for the same attributes using the human ratings of the OMI dataset. I then measured the similarity of C_m and C_h based on the normalized Frobenius inner product $\text{F}_{m,h}$. I used a one-way ANOVA to test for differences in $\text{F}_{m,h}$ between the 80M, 400M, and 2B models. I used a paired-samples t-test to conduct post hoc comparisons.

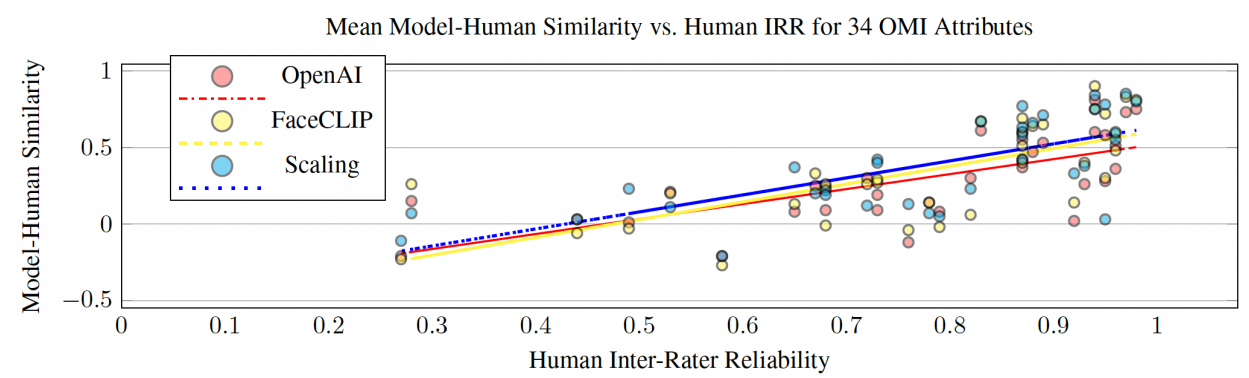


Figure 4.3: The similarity of CLIP bias to human bias is strongly correlated with human IRR, indicating that the societal consistency of a bias plays a significant role in whether a model learns it during semi-supervised pretraining.

Peterson et al. (2022) [304] study the structure of facial impressions by computing the correlation matrix of OMI ratings and qualitatively examining its hierarchical structure. The present research also qualitatively compares the structure of the OMI correlation matrix to the hierarchically clustered correlation matrix of a CLIP-ViT-L-14 trained on LAION-2B for 13-billion total samples. If the structure of OMI biases is similar to CLIP, I expect to observe similar attribute clusters at higher levels, with differences emerging in leaf nodes.

4.5.4 Generative Text-to-Image Models

Finally, I extended the analysis to SDXL-Turbo, SD2, and Runway SD1.5. I first studied the similarity of these models' representations to human attribute ratings. To do so, I generated N=25 images for each attribute's positive pole and 25 for its negative pole. I adjusted prompts for the models to "a realistic portrait photo of someone who is *attribute*," because image generators may create cartoonish images that do not center the face. I extracted embeddings for these images and computed their projection products g^{a+} and g^{a-}. If generative text-to-image models reflect human-like facial impression biases, I expect images generated from a positive prompt to have positive projections, and images from a negative

prompt to have negative projections. I thus frame model-human similarity as a classification problem, wherein images generated from the positive pole prompt receives a label of 1, and from the negative pole prompt receive a label of 0. The OMI subspace is positioned as a classifier, which predicts 1 where an image vector's projection product is positive, and 0 where it is negative. I report Recall, Precision, and F1 Score. I validated this approach using the Outdoors attribute, a control group for measuring validity in the OMI dataset, obtaining F1=.94 for SDXL-Turbo

I then study social bias in Stable Diffusion XL-Turbo by projecting the positive prompt images generated for the White and Black attributes onto all 34 of the OMI attribute subspaces. For each subspace, I compute the White-Black differential bias by obtaining an effect size (Cohen's d) between the projection products $\mathbf{g}^{\mathbf{White}^+}$ and $\mathbf{g}^{\mathbf{Black}^+}$, then measure statistical significance using a paired samples t-test.

4.6 Results

Results indicate that 1) CLIP models exhibit facial impression biases; 2) Human IRR is a significant predictor of which biases are learned; 3) models trained on larger datasets exhibit emergence of subjective facial impression biases, and more human-like associations among impressions; and 4) text-to-image generators exhibit facial impression biases and undesirable social biases associating preferred attributes with images of White individuals.

4.6.1 CLIP Models Reflect Human Biases

As shown in Figure 4.1, OpenAI, FaceCLIP, and Scaling CLIP models exhibit human-like facial impression biases. Relatively objective attributes like age, hair-color, and happiness exhibit high model-human similarity, as do many socially constructed attributes such as gender and cuteness. Notable exceptions include Black, White, and skin color attributes, which fall short of expectations based on Human IRR, as visualized in Figure 4.3. Model-human similarities of traits like Trustworthiness, Electability, and Intelligence are significant but lower, consistent with the lower IRR of these attributes. That CLIP learns these biases

Model-Human Similarity of Facial Impression Bias in CLIP Models by Pretraining Dataset Size									
Measurement	Mean (Std)			Max			Cohen's d		
Attribute	2b	400m	80m	2b	400m	80m	2b-80m	400m-80m	2b-400m
Happy	.88 (0.02)	.86 (0.03)	.76 (0.06)	.91	.90	.84	1.54*	1.36*	0.73
Gender	.85 (0.02)	.87 (0.02)	.83 (0.04)	.87	.91	.89	0.61	1.08*	−0.84*
Age	.83 (0.08)	.84 (0.05)	.73 (0.11)	.93	.90	.90	0.94	1.03*	−0.04
Asian	.82 (0.03)	.81 (0.05)	.73 (0.06)	.85	.86	.79	1.40*	1.13*	0.43
Hispanic	.82 (0.02)	.80 (0.02)	.67 (0.07)	.85	.84	.78	1.62*	1.50*	1.07*
Outdoors	.81 (0.01)	.79 (0.04)	.65 (0.10)	.82	.84	.75	1.51*	1.39*	0.57
Pacific Islander	.76 (0.03)	.74 (0.04)	.62 (0.11)	.81	.81	.77	1.28*	1.15*	0.50
Middle Eastern	.72 (0.02)	.68 (0.05)	.57 (0.11)	.75	.78	.71	1.43*	1.17*	0.88
Native American	.68 (0.06)	.67 (0.06)	.53 (0.13)	.78	.78	.77	1.20*	1.15*	0.14
Weight	.68 (0.03)	.69 (0.03)	.63 (0.05)	.72	.72	.67	1.05	1.24*	−0.40
Hair-Color	.66 (0.07)	.61 (0.09)	.50 (0.12)	.76	.70	.67	1.31*	0.92	0.67*
Cute	.65 (0.10)	.67 (0.08)	.41 (0.15)	.78	.76	.56	1.35*	1.46*	−0.31
Long-Haired	.63 (0.10)	.63 (0.11)	.43 (0.16)	.76	.75	.67	1.20*	1.16*	0.06
Gay	.49 (0.07)	.41 (0.06)	.30 (0.11)	.57	.51	.42	1.46*	1.09*	1.14*
Attractive	.48 (0.08)	.50 (0.11)	.28 (0.21)	.60	.65	.57	1.11*	1.11*	−0.16
Electable	.47 (0.22)	.50 (0.11)	.30 (0.25)	.68	.65	.60	0.72	0.94	−0.16
Smart	.42 (0.15)	.45 (0.14)	.25 (0.11)	.62	.59	.50	1.09*	1.23*	−0.20
Black	.41 (0.09)	.35 (0.11)	.37 (0.12)	.51	.49	.51	0.46	−0.10	0.59
Smug	.38 (0.11)	.19 (0.20)	.07 (0.15)	.50	.45	.25	1.50*	0.63	1.02*
Trustworthy	.36 (0.18)	.04 (0.18)	-.06 (0.15)	.59	.35	.15	1.56*	0.60*	1.33*
Skin-Color	.36 (0.14)	.37 (0.13)	.28 (0.11)	.59	.58	.43	0.58	0.70	−0.11
Outgoing	.33 (0.13)	.21 (0.17)	.16 (0.16)	.52	.35	.42	1.02*	0.31	0.74
Privileged	.26 (0.15)	.21 (0.12)	.05 (0.07)	.49	.44	.16	1.35*	1.31*	0.33
Godly	.20 (0.09)	.26 (0.17)	.25 (0.10)	.34	.51	.42	−0.57	0.01	−0.42
Liberal	.14 (0.15)	.22 (0.15)	.23 (0.19)	.32	.45	.50	−0.49	−0.05	−0.51
Typical	.13 (0.10)	.11 (0.16)	.05 (0.13)	.28	.28	.20	0.72	0.46	0.16
Dorky	.13 (0.23)	.17 (0.22)	.05 (0.21)	.38	.52	.33	0.37	0.52	−0.15
Familiar	.12 (0.12)	.04 (0.08)	-.00 (0.14)	.32	.14	.16	0.90*	0.42	0.75*
Well-Groomed	.12 (0.26)	-.04 (0.21)	.11 (0.18)	.43	.44	.31	0.03	−0.71	0.62
Dominant	.11 (0.23)	.06 (0.18)	.00 (0.29)	.47	.34	.29	0.41	0.25	0.23
Memorable	.08 (0.11)	-.05 (0.15)	.04 (0.12)	.24	.13	.18	0.35	−0.67*	0.92*
White	.00 (0.28)	.12 (0.21)	-.03 (0.20)	.45	.45	.26	0.12	0.67	−0.46
Looks-Like-You	-.13 (0.14)	-.10 (0.14)	-.11 (0.16)	.15	.06	.11	−0.17	0.08	−0.26
Alert	-.29 (0.10)	-.17 (0.10)	-.14 (0.09)	-.16	-.03	.05	−1.25*	−0.31	−1.05*

Table 4.1: 17 of 34 of attributes exhibit significant differences and large effect sizes between model groups.

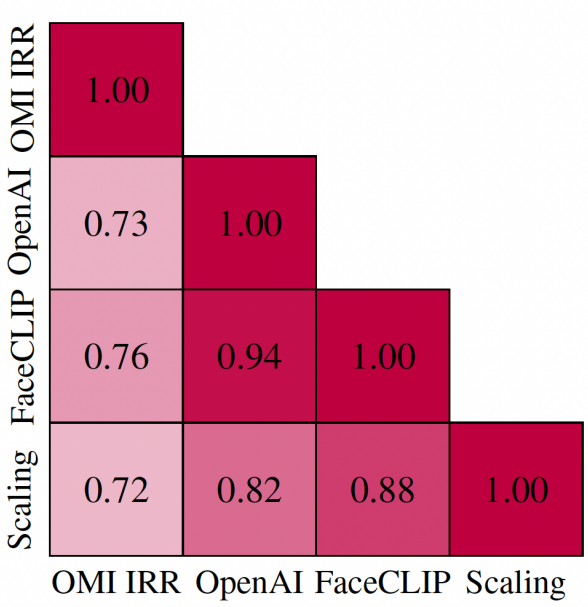


Figure 4.4: CLIP models exhibit significant Spearman's ρ between Mean Model-Human Similarity and OMI IRR.

at all is noteworthy: to my knowledge, this is the first research to document unobservable facial impression biases learned by a semi-supervised vision-language model (rather than a supervised model of facial impressions) consistent with human societal biases.

As described in Figure 4.4, all three families of models exhibit strong correlations ranging from .72 to .76 between the mean model-human similarity of a trait and its Human IRR. Coefficients are larger between OpenAI and FaceCLIP models than with Scaling CLIP models, likely a result of FaceCLIP post-training from OpenAI base models.

4.6.2 Dataset Scale

A multiple linear regression finds that only Human IRR and Dataset Size are statistically significant predictors of model-human similarity. As described in Table 4.2, Human IRR plays the larger role of the two independent variables, as evidenced by a much larger coefficient and t-value. Total Training Samples, Image Parameters, and Text Parameters are not statistically

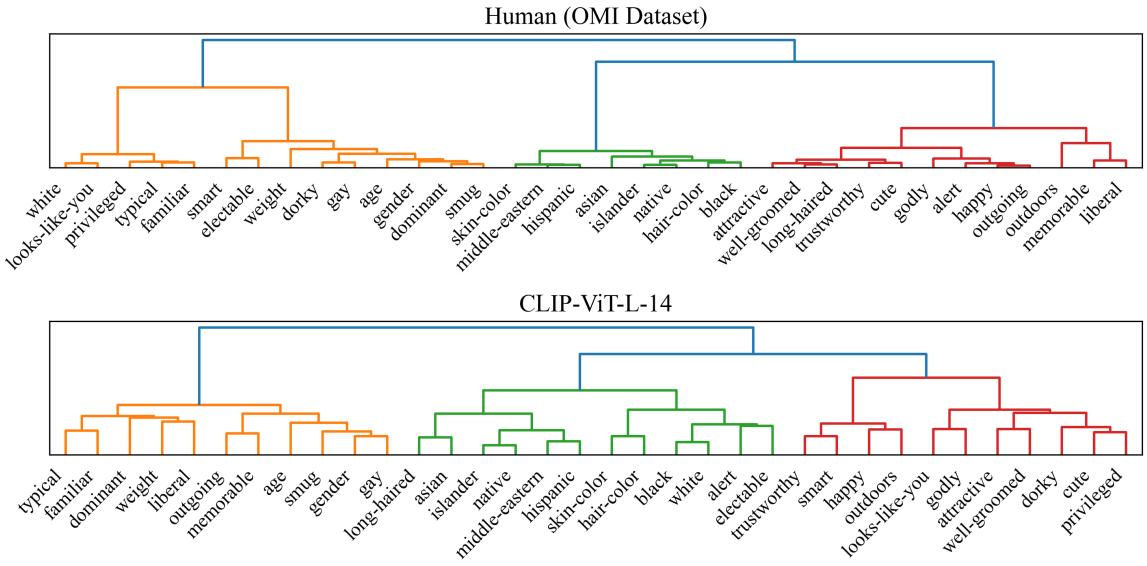


Figure 4.5: The structure of facial impression biases in CLIP-ViT-L-14 mirrors that of human facial impression biases quantified in the OMI dataset. Clusters related to ethnicity emerge in each, as do clusters grouping gender, sexuality, and smugness.

Statistic			
Adj. R^2	.420		
F-Statistic	134.0		
N	918		
DoF Residuals	912		
DoF Model	5		
Ind. Variable	Coef	t	$p < \|t\|$
Human IRR	1.1204	25.470	.001
Dataset Size	.0877	4.484	.001
Total Samples	−.0132	−.599	.549
Image Params	−.5615	−.150	.881
Text Params	.7898	.144	.885
Constant	−.7542	−.434	.665

Table 4.2: Fitting a linear regression to model-human similarity coefficients reveals that Human IRR and Dataset Scale are significant predictors a bias will be learned by a CLIP model.

significant predictors of facial impression bias in vision-language models.

Table 4.1 describes the mean (with standard deviation) and maximum model-human similarity for each dataset size, and it reports Cohen's d between the groups of models trained on the three dataset sizes. Effect sizes obtained between the 400M and the 80M level are large and statistically significant for 17 of 34 attributes, with large absolute differences between attribute means, such as .41 for Cute at the 80M level vs. .67 at the 400M level. For most attributes, comparisons are not statistically significant between the 2B and 400M levels, though they may return small or medium effect sizes. A notable exception emerges for several unobservable attributes, including Trustworthiness and Sexuality, which exhibit large effect sizes and statistically significant differences between the 2B level and the 400M level. While observable attributes such as Happpiness reflect human ratings well at the 80M level, and more subjective but still visually observable attributes like Cute are reflected consistently at the 400M level, it is not until the 2B level that CLIP models reflect subjective and

visually *unobservable* attributes such as Trustworthiness. The results indicate that increases in the scale of the pretraining data have more significant effects for learning subtle societal biases reflecting attributes with *lower* IRR. Models trained on additional data approximate a distribution that more closely reflects the perceptions of society as a whole, learning to use the biased visual heuristics present in the human-authored captions in the pretraining data, even for unobservable attributes.

4.6.3 Structure of Facial Impressions

Results indicate that dataset size impacts the extent to which the *structure* of facial impression bias in CLIP reflects the structure of the facial impression bias in humans. Figure 4.5 visualizes the hierarchical similarities between the attribute cross-correlation matrix for CLIP-ViT-L-14 (the most commonly used CLIP model as of this writing) and the OMI attribute cross-correlation matrix. The most salient similarities between the two include a cluster of correlated racial and ethnic identities, such as Hispanic, Middle-Eastern, Native American, and Pacific Islander, as well as a cluster grouping together the Smugness, Gender, Sexuality, and Age attributes. There are also differences between the model and human ratings: while Trustworthy is correlated with Cute in OMI, it is correlated with Smart and Happy in CLIP. Similarities between CLIP attribute correlations and OMI attribute correlations are more evident at the higher levels of the hierarchy, with differences appearing toward the leaf nodes.

A one-way ANOVA provides evidence that similarities in the structure of facial impression biases increase for models trained on larger datasets, with $F(2) = 15.71$, $p < .001$ after Bonferonni correction, demonstrating statistically significant differences in $\mathrm{F}_{m,h}$ between the Scaling-2B, Scaling-400M, and Scaling-80M models. One can observe statistically significant differences and large effect sizes both between the 80M and 400M levels and between the 80M and 2B levels with post-hoc tests. Figure 4.6 illustrates the differences among the three model groups, showing that the magnitude of difference is greater between the 80M level and the 400M level than between the 400M level and the 2B level.

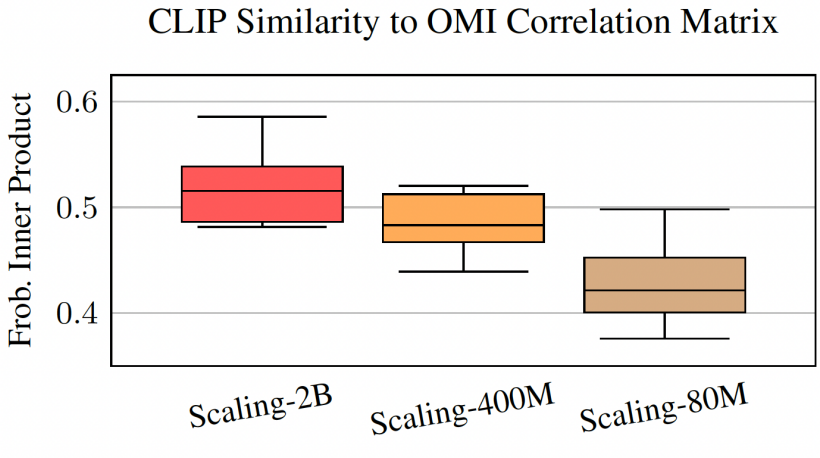


Figure 4.6: Scaling-2B CLIP models exhibit the greatest structural similarity to human facial impression biases.

4.6.4 Generative Text-to-Image Models

As shown in Table 4.3, one can observe more variance in text-to-image generator F1 scores than in CLIP model-human similarities. SDXL has the most human-like associations, with Spearman's $\rho = .60, p < .001$ between Human IRR and SDXL F1 scores. Runway-SD1.5 is also correlated with IRR, with $\rho = .50, p < .01$, while SD2 is not significantly correlated with IRR, with $\rho = .32, p = .06$. Notably, the Attractive attribute has the highest F1 score for SDXL, whereas it is the 16th most human-similar attribute in CLIP models, suggesting the importance of representing beauty for user-facing image generators, which often undergo additional training to better reflect user aesthetic preferences [345]. With the exception of Liberal, unobservable traits rank in the bottom half of F1 scores for SDXL, and Trustworthy is lowest of any trait. Though SD2 and Runway-SD1.5 are more human-like than SDXL for trustworthiness, the results suggest that exploiting biased heuristics may be more straightforward for a classifier like CLIP.

Images generated by Stable Diffusion XL-Turbo also bear signs of racial bias: as shown in Figure 4.7, one observes statistically significant differences indicating that generated images

Attribute F1 Scores by Stable Diffusion Model			
Attribute	SDXL	SD2	Runway-SD1.5
attractive	0.98	0.7	0.65
outdoors	0.94	0.67	0.60
well-groomed	0.91	0.66	0.59
hair-color	0.83	0.79	0.77
weight	0.76	0.57	0.37
long-haired	0.76	0.67	0.65
black	0.75	0.76	0.76
white	0.74	0.43	0.53
asian	0.70	0.63	0.69
middle-eastern	0.69	0.70	0.64
cute	0.69	0.58	0.55
happy	0.68	0.79	0.77
islander	0.68	0.66	0.69
age	0.68	0.68	0.72
liberal	0.68	0.60	0.62
skin-color	0.67	0.68	0.61
alert	0.67	0.57	0.59
gender	0.66	0.64	0.62
smart	0.65	0.57	0.71
dominant	0.65	0.68	0.67
hispanic	0.65	0.66	0.69
native	0.64	0.69	0.68
electable	0.61	0.65	0.60
dorky	0.53	0.60	0.60
looks-like-you	0.53	0.68	0.49
smug	0.51	0.58	0.57
memorable	0.48	0.66	0.59
privileged	0.47	0.60	0.58
gay	0.37	0.56	0.40
godly	0.31	0.65	0.59
typical	0.29	0.63	0.63
familiar	0.27	0.68	0.51
outgoing	0.24	0.66	0.71
trustworthy	0.12	0.51	0.64

Table 4.3: SDXL-Turbo reflects human facial impression biases, with Spearman's $\rho = .60$ between IRR and SDXL F1.

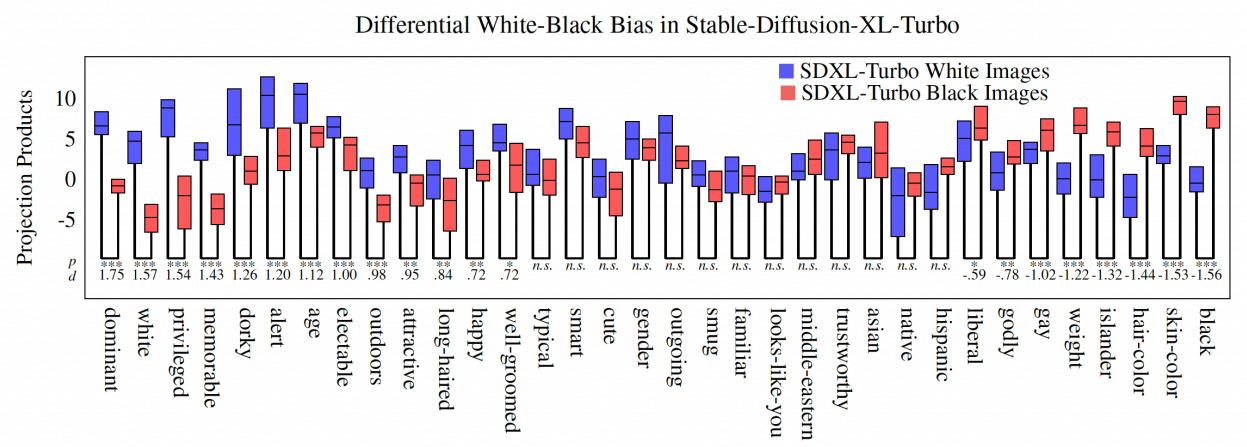


Figure 4.7: Stable Diffusion XL-Turbo exhibits differential White-Black biases, projecting White individuals as more dominant, electable, and attractive. I plot effect size and significance (* $<$ .05, ** $<$ .01, *** $<$ .001) below significant comparisons.

of White individuals are more likely to be perceived as dominant, privileged, memorable, attractive, electable, and happy than images of Black individuals, which are more likely to be perceived as more liberal and heavy (vs. thin). Note that many of these relationships are *not* observed in the OMI correlation clusters seen in Figure 4.5, indicating that they originate not with the OMI dataset but with the text-to-image model.

4.7 Discussion

Results make clear the inter-connection of visual perception in AI with the human social world: where a facial impression bias is more consistently shared among humans, CLIP is also more likely to learn it. Facial impression biases have notable consequences in professional and civic life [374], and prove difficult to dislodge even after intervention [186]. While CLIP may serve as a tool for studying such biases, it may also reinforce or amplify these biases in society, especially given their presence in user-facing image generators.

4.7.1 Scale and Bias

Training on larger datasets results in emergent and amplified facial impression biases. CLIP models exhibit more human-like biases related to trustworthiness and sexuality when trained on LAION-2B, and nearly every OMI attribute increases in model-human similarity between the 80M level and the 400M level. That model parameterization plays no detectable role in facial impression bias underlines that what CLIP models have learned is a biased visual heuristic reflected in the training data, not a more precise representation of an objectively detectable attribute. Greater attention to the characteristics of the training data is also advisable when pretraining CLIP systems for use in downstream applications without fine-tuning, or for supervising other models, including text to image generators using CLIP as a text encoder.

4.7.2 Ramifications of Human-like Models

Amid the excitement over vision-language models like GPT-4 that convincingly imitate aspects of human intelligence [61], the emergence of subtle biases in multimodal models trained on the largest datasets elicits a corrective question: is more "human-like" always better? Approximating the distribution of societal associations present in the pretraining dataset as completely as possible may be useful for providing a more general-purpose model. However, as zero-shot vision-language models continue to become more accessible to the general public, including via conversational interfaces that mimic text-based interaction with a human being [221, 228], monitoring for subtle emergent biases in need of mitigation is likely to become a more pressing concern. Humans interacting with fluent, mostly debiased models may be less skeptical of *subtle* reflections of societal bias than of the more blatant misrepresentations of demographic groups in previous models.

4.7.3 Implications for Computational Social Science

The present work is also notable in its consequences for computational social science. By my estimation, Peterson et al. (2022) [304] spent tens of thousands of dollars to collect the human subject data needed to learn a supervised model of facial impressions; CLIP produces a model of facial impressions as a side effect of pretraining. While CLIP provides a less precise model of these biases than the supervised counterpart of Peterson et al. (2022) [304], this research nonetheless suggests that CLIP models might play a role similar to static word embeddings, which social scientists now employ in computational studies of human attitudes, including in research that generalizes the findings of human subjects experiments [252, 67], or quantifies shifts in human attitudes over decades [52, 142]. That CLIP models reflect facial impression bias suggests that they could model other complex sociocultural phenomena not observable via text embeddings alone.

4.7.4 Limitations

While participants in the study of Peterson et al. (2022) [304] were reflective of U.S. demographics as a whole, this also means that a majority of perceivers identified as White, as is clear from the correlation of White and Looks-Like-You OMI attributes in Figure 4.5. Such a demographic skew may render attribute ratings sensitive to correlated biases, given that prior work observes a relationship between social bias and face impressions [450]. In addition, while I adopt now-standard prompts specified by OpenAI when introducing CLIP Radford et al. (2021) [321], significantly changing these prompts may induce variance in the results.

4.8 Conclusion

The present work demonstrates that training general-purpose vision-language AI on ever-larger datasets produces novel epistemic risks, even as it improves the accuracy of these models on benchmarks and traditional machine learning tasks. Such models may serve as useful tools for social science, potentially illuminating factors that contribute to human bias.

However, the epistemic risk induced by these biases in vision-language AI also renders fraught these models' real-world use, as they may subtly reinforce existing inequities, including in an online environment increasingly mediated by AI.

Chapter 5

DESIGNING FOR VERIFICATION: AN APPROACH TO THE EPISTEMIC RISKS OF AI IN HIGH-STAKES KNOWLEDGE WORK

5.1 Preface

Epistemic risk that prevents general-purpose AI from reaching its full potential does not necessarily prevent its responsible use today. In the case of facial impression bias, BeMyAI can still provide a remarkable interface to most of the world for low-vision users, even when it is programmatically prevented from describing images of people. While developing effective technical solutions to epistemic risk remains an important research direction, and one I have explored in my collaborations [454], an equally important direction explores how general-purpose AI systems can be responsibly deployed *in the presence of inevitable epistemic risk*. In this third study, I create a novel dimension in the design space of generative models [255] intended to reduce epistemic risk in human-AI collaboration. To do so, I specifically sought out the input of professional fact-checkers, a group in the process of reckoning with the transformations of generative AI for their work. The findings of this chapter support **Thesis Statement 3**: *Design that centers the verification of information can mitigate epistemic risk when using generative and general-purpose AI in the production of knowledge.*

5.2 Introduction

A research report issued by OpenAI in March 2023 [127], days after the release of its flagship GPT-4 model, contended that generative pretrained transformers (GPTs) are general purpose technologies, technologies with the potential to reshape not an individual profession but an entire economy. Unlike many previous general purpose technologies, the authors asserted that

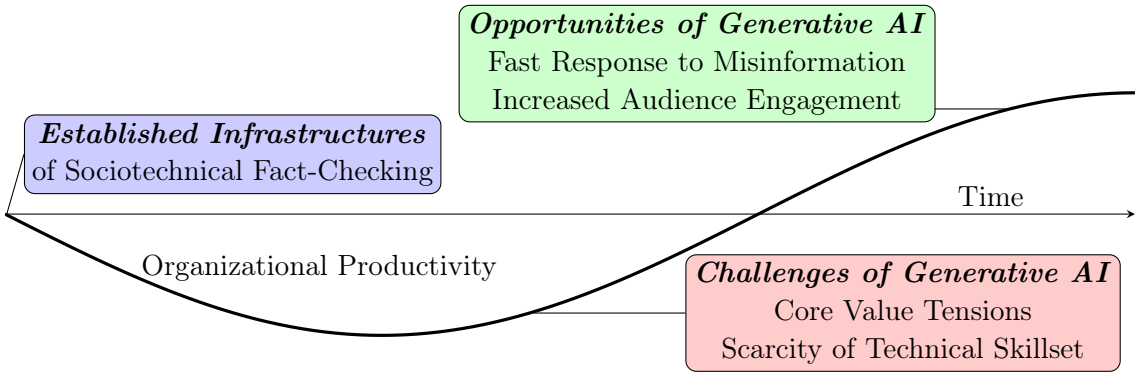


Figure 5.1: Incorporating a society-altering technology like generative AI into sociotechnical fact-checking work requires "intangible" investments [60] (new processes and skills) to realize its potential without deprioritizing the values of fact-checking or displacing the role of human experts.

generative AI will impact primarily professions with a higher barrier to entry, those requiring more education and experience to carry out. Among the professions estimated by an OpenAI model as "fully exposed" to transformation by generative AI, defined as reducing by at least 50% the time needed to complete the tasks of an occupation, was "News Analysts, Reporters, and Journalists" [127]. Yet these professions have outsized epistemic effects on society [126], as they remain the primary means for producing knowledge claims and for critically assessing sources and information [125], thus ensuring the integrity of the online information space. If generative AI is to reshape such roles, understanding how it might do so – and where to draw the boundaries – is crucial to ensure the health of the information ecosystem.

In this work, I study the impact of generative AI in fact-checking, a profession that specializes in determining the reliability of information disseminated through traditional and social media, undertaken at publishing houses and independent organizations around the world. Fact-checking is a complex sociotechnical process, involving human judgment exercised in conjunction with AI-based tools to observe misinforming claims and narratives as they spread [160]. While most fact-checking organizations necessarily embrace technological tools,

they are skeptical of technologies that promise to automate large parts of the fact-checking process, and deprioritize or displace human expertise [195]. Understanding perspectives of key stakeholders at fact-checking organizations is thus important to facilitate adoption of a technology that could help respond efficiently to misinformation, while prioritizing the role of human expertise. I address two research questions:

1. **RQ1: Opportunities of Generative AI in Fact-Checking:** What opportunities do fact-checking organizations see in generative AI? How are organizations presently using generative AI, and how do they envision using it?

2. **RQ2: Challenges and Limitations of Generative AI in Fact-Checking:** What challenges do fact-checkers see in using generative AI to support their work? What prevents them from further incorporating generative models?

To address these questions, I interviewed N=38 participants at 29 fact-checking organizations in a range of roles, from investigation, to management, to engineering. I captured diverse, global perspectives from participants located across 19 countries and six continents. Interviews provided detailed accounts of where fact-checkers envisioned using generative AI, and concrete examples of applications in use or in development. Participants also shared barriers to adopting generative AI, ranging from technical limitations to value misalignments. Figure 5.1 draws on organizational research [60] to illustrate the investment to overcome these challenges and realize the benefits envisioned by participants. I make four contributions:

- **Enumerating Opportunities and Limitations of Generative AI in Fact-Checking:** I describe the opportunities for generative AI in five fact-checking infrastructures (Editing, Investigation, Audience Management, Technology, and Advocacy), and adopt the Technology-Organization-Environment framework [313] to describe challenges.

- **Designing for Verification:** I propose a novel dimension in the design space for generative models that centers Verification, or ensuring the veracity of content. I describe

this dimension with a 2x2 matrix, with the Producer of content on the X axis, and its Verifier on the Y axis, and discuss its use beyond fact-checking in high stakes domains.

- **Mapping Value Tensions:** Using the principles of the International Fact Checking Network (IFCN) [311] as a basis for the sociotechnical values of fact-checking, I describe value tensions between fact-checking, which centers transparency and reliability, and generative AI, a technology exhibiting unpredictable and often unreliable behavior.

- **Defining a Research Agenda:** I propose nine directions for fairness, accountability, and transparency researchers to develop technologies, designs, and approaches supporting responsible use of generative AI in fact-checking.

5.3 Methods

5.3.1 Participant Recruitment

I conducted an interview study with N=38 employees of fact-checking organizations or teams in publication houses, with experience in their current role ranging from 1 year to 18 years. As shown in Table 5.1, I recruited from a total of 29 fact-checking organizations using purposive sampling and snowball sampling [129, 261], first reaching out to potential participants by sending an email with my senior co-author to advertise the study to the listserv of the IFCN. This resulted in three interviews with six participants working at three organizations. The Community Manager of the IFCN then provided contact information for six potential participants for the study, to whom I reached out. This resulted in three interviews with three participants at three organizations. I next utilized a list of 23 fact-checkers known to my senior co-author, who has maintained a long-term relationship with the global fact-checking community. This resulted in five interviews with six participants at five organizations. Finally, I sent cold emails to 60 IFCN signatory organizations, explaining my interest in an interview and how I found their contact information. I recruited in this way not only to increase the number of participants in the study, but also to increase the study's global reach, as I emailed

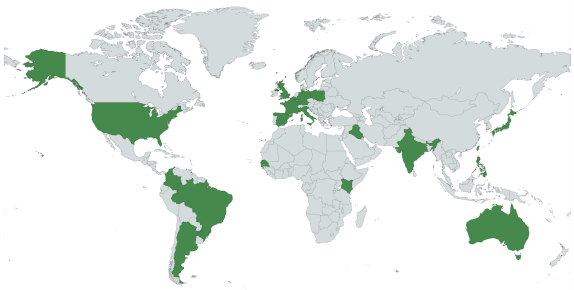
Continents (6) | Countries (19)	Fact-checking Organizations (29)
	Australian Associated Press [315], Agence France-Presse [316], Africa Check [77], Aos Fatos [131], Chequeado [83], Code for Africa [135], ColombiaCheck [92], Der Spiegel [369], Factly [130], India Today [395], Infoveritas [183], Lead Stories [375], Litmus [227], logically.ai [229], Maldita.es [236], Meedan [241], MindaNews [249], Newtral [264], Pagella Politica [307], PolitiFact [308], Pravda [314], Rappler [333], RMIT FactLab CrossCheck [96], Science Feedback [132], Taiwan FactCheck Center [73], Tech4Peace [388], The Quint [320], Thomson Reuters [337], Univision El Detector [105]

Table 5.1: Participants were recruited from 6 continents, 19 countries, and 29 fact-checking organizations.

primarily organizations in developing countries and the global south. This strategy resulted in 14 interviews with 18 participants at 14 organizations across five continents. I employed snowball sampling when participants offered to connect me with a participant well-suited to the study, and reached out via email. This resulted in five interviews with five individuals working at five organizations.

5.3.2 Interview Protocol

I created a semi-structured interview protocol that posed general questions regarding the use and impact of generative AI in fact-checking. I began interviews by asking participants to tell me about their background, including their position, experience in fact-checking, and

familiarity with generative AI. I asked about their company's background, including the size and technical experience of the fact-checking team, and how long the company had been performing fact-checking work. I then explicitly posed the primary research questions of the study, asking participants to characterize 1) how they used generative AI in their work; 2) opportunities for using generative AI in fact-checking; 3) challenges and limitations of using generative AI in fact-checking; and 4) how researchers could design generative technologies that better support fact-checkers. I asked participants to clarify, discuss, and expand upon responses to better understand their perspectives. I also asked follow up questions where appropriate about several specific topics, including the use of corporate vs. open source AI; modalities (text, image, etc.) of misinformation they use generative AI to handle; use of generative AI to handle narratives; guidelines for using generative AI in their organization; and impacts the participant witnessed generative AI having. I submitted the interview protocol as part of the supporting materials to my University's Institutional Review Board.

5.3.3 Interview Process

I conducted 30 interviews between October 2023 and January 2024. Multiple participants attended seven interviews, with one participant typically a manager, and the other(s) involved in technology or investigation. In one case, I interviewed two managers who passed follow-up questions to the engineering team, forwarding their responses by email. Interviews lasted 30 to 90 minutes, averaging approximately 45 minutes. Interviews were conducted solely in English. I accommodated the request of one participant to send answers by email because they preferred writing over speaking in English. The participant then also met with me over Zoom for 20 minutes. Participants who used generative AI sometimes shared their screen and displayed the interfaces used with these technologies. One participant shared a Jupyter notebook showing their use of AI in a data science pipeline. Other participants linked me to Github pages or company technical reports. I did not offer to compensate participants to prevent feelings of coercion.

5.3.4 Data Analysis

After transcribing the interviews, they were deductively coded according to four categories tied closely to the research questions: Present Use of Generative AI in Fact-Checking; Opportunities to Use Generative AI in Fact-Checking; Challenges and Limitations to Using Generative AI in Fact-Checking; and Ways Computational Research in Fairness and Transparency Can Support Fact-Checking. They were then inductively coded within each deductive category.

My co-author and I first coded four interview transcripts, after which I created a codebook that included inductively derived codes organized within the deductive categories. The codebook consisted of the names of codes, explanations of the codes, and the associated participant quotes. I shared the codebook, and my co-author offered feedback and suggestions, after which we met to discuss the codes and revised or removed codes on which they could not reach agreement. We then coded four additional transcripts at a time, updating the codebook after each round with more precise definitions and additional context from participant quotes. Next, we followed a thematic analysis process [90, 55] to generate themes that described the findings and addressed the research questions. Specifically, we reviewed the codes and their associated participant quotes, drafted memos describing proposed themes, met to discuss the proposed themes, and converged on a set of final themes on which agreement could be reached. These themes form the basis of the Findings section.

5.4 Findings: Opportunities and Challenges of Generative AI in Fact-Checking

5.4.1 RQ1: Opportunities of Generative AI

I found during thematic analysis that the technologies used and envisioned by fact-checking organizations depend largely on the organizational infrastructures into which they would be integrated. These infrastructures accorded to a large degree with the work of Juneja and Mitra (2022) [195], which described the sociotechnical work of editors; fact-checkers; social media managers; and long-term advocators. Drawing on these roles, I organize the findings according to the following divisions of organizational infrastructure, as shown in Fig. 5.2:

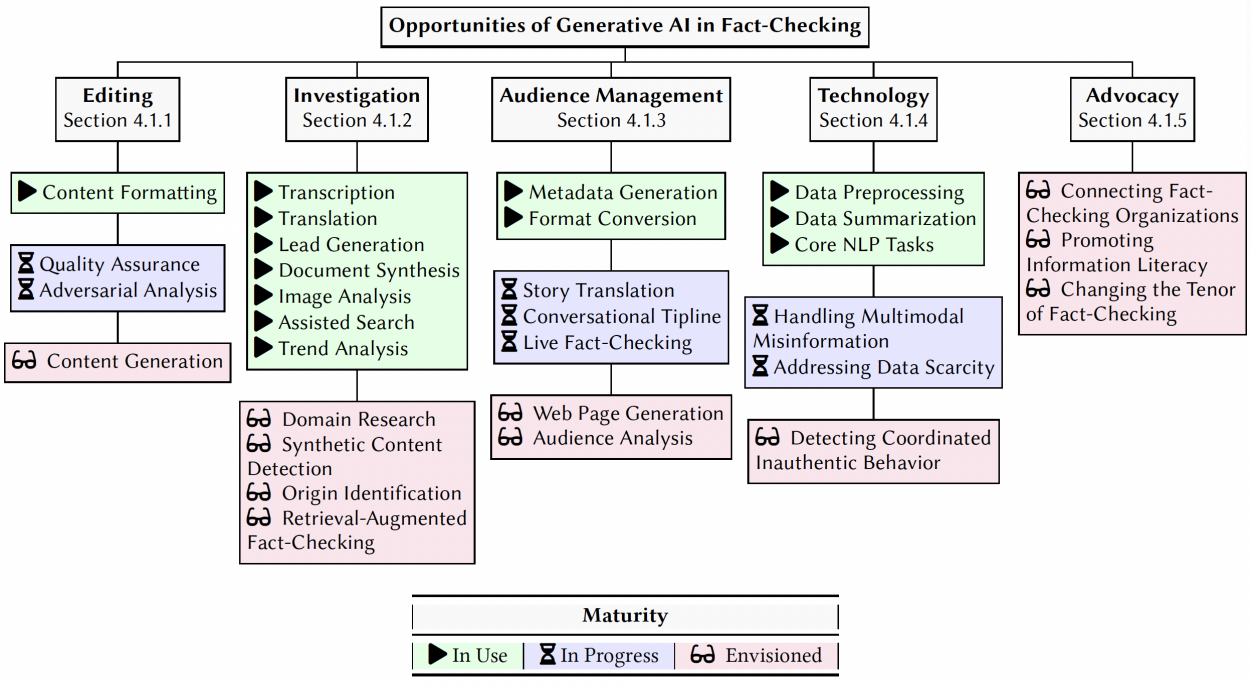


Figure 5.2: A description of In Use, In Progress, and Envisioned generative technologies grouped according to five fact-checking infrastructures.

Editing, Investigation, Audience Management, Advocacy, and Technology, the last of which is new to this work but necessary to describe the impact of generative AI on the work of software developers and data scientists who build and maintain the data science pipelines employed by fact-checking organizations.

When describing a technology, I also identify its status according to three levels of maturity, denoted using icons:

- ▶ **In Use:** Technologies presently in use by participants, denoted with a rightward arrow to evoke a "Play" symbol.
- ⧗ **In Progress:** Technologies undergoing prototyping, testing, betas, or development by participants, denoted with an hourglass symbol to communicate that some time remains before these technologies will be implemented.
- 👓 **Envisioned:** Technologies envisioned but not implemented or prototyped by participants, denoted with eyeglasses to communicate that these technologies are further away and not yet in development.

Where fact-checking organizations reported achieving differing levels of maturity for a technology, I describe the most mature, and make note if this level of maturity has not been achieved by most other fact-checking organizations.

Editing

Editing ensures fact-checking content is engaging, approachable, and error-free, and spans from the beginning of a fact-check to publication, as editors are often involved in deciding which claims are worthy of a fact-check [195]. Participants in the study who were involved in Editing usually reported managing small teams.

Many participants described using generative AI to refine and restructure fact-checks and internal reports, which then undergo human review. P3 reported using "premium" ChatGPT

for ▶ **Content Formatting**—to edit and refine written reports, and to restructure "dense content" into an approachable format for readers. P18 noted using ChatGPT to help in "brushing up text." Several participants described in-development applications of generative AI for systematic ⧗ **Quality Assurance**, to prevent cosmetic and substantive editorial errors. P35 reported using ChatGPT to highlight grammatical and factual errors, tasks for which they normally use Grammarly [152]. P27 developed an app to address such errors:

> We began with simple mistakes, geographical errors, or misspellings [...] I made a little Shiny app out of [ChatGPT] and showed it around, and people were really like [...] I can really see how this helps us. [...] And my aim is that we don't have these kind of simple mistakes anymore [...] this would be a huge achievement because simple mistakes are trivial on the one hand, but on the other hand [...] it's really important for the sense of quality the reader has and for trust.

P3 expanded on this view and envisioned generative AI providing ⧗ **Adversarial Analysis** before publishing a fact-check, referencing a strategy in development at a Sudanese newsroom: "They use generative AI to actually give it an article that has been written, and ask the model to actually tell us whether there are any assumptions that have been met, that are inaccurate or incorrect." Finally, P8 envisioned fine-tuning a generative model on verified fact-checking articles, and using it for ∞ **Content Generation** at the beginning of the composition process, improving "fact-checking output [by] pre-writing those fact-check articles," provided that the content did not require deep scientific knowledge of a topic, and remained subject to human review. While some participants envisioned using generative AI for content generation, many expressed deep discomfort with generative AI writing fact-checks, a finding explored in sections 5.4.2 and 5.5.1.

Investigation

Investigation refers to the process of assessing the accuracy of potentially misinforming claims, and involves monitoring online sources for misinforming content; gathering verified sources to

substantiate or refute claims; and writing a fact-check or internal report [195]. 19 of the 38 participants described working primarily in Investigation.

Participants described many ongoing and envisioned uses of generative AI to perform tasks related to investigation and research assistance. As P25 noted, generative AI is used in common and "taken-for-granted" tasks like ▶ **Transcription**, which save time and money for fact-checkers. P20 highlighted the use of generative AI for ▶ **Translation** of internet content in need of investigation, such as "many translations from Ukraine" due to misinformation related to the Russia-Ukraine war. Participants also described adopting generative AI to directly support investigations. P11 noted that, while they never use ChatGPT for writing fact-checks, they use it for ▶ **Lead Generation**, "trying to generate ideas for stories." P28 used ChatGPT for ▶ **Document Synthesis**, to save time by organizing research notes and summarizing text from web pages. P19 used GPT-4 for ▶ **Image Analysis**, substituting the model for reverse image search in some cases, noting they can "ask it where the photo was taken, and sometimes we [get the] correct answer," or useful hints for continuing the search. P25 fine-tuned GPT-3.5-Turbo to perform ▶ **Assisted Search**, generating custom Google search queries, often in a language not spoken by the fact-checker, noting that such a task would "take me hours to do and I still might miss some of the terms." P14 reported using ChatGPT for ▶ **Trend Analysis**, to keep abreast of media produced by websites known to produce misinforming content: "We take the top 200 headlines from the last 24 hours from those sites [...] and run them through ChatGPT, asking it to summarize the main narratives [...] and extract the names of people, places, entities [...] and then send that to me by email. So every six hours, [...] we get an email."

Participants envisioned technologies to increase their expertise and verify novel content. P29 noted fact-checkers need to "become a little mini expert in a certain specific topic," and envisioned a technology for **Domain Research**, summarizing and collating literature for review. P24 envisioned using generative AI for **Synthetic Content Detection**, identifying content produced by AI. P24 described this as their "holy grail" given recent increases in synthetic content and the difficulty of fact-checking it. P29 envisioned a tool

for **Origin Identification**, scanning the internet for the first occurrence of content, bringing fact-checkers "closer to the verification." Finally, P26 envisioned a **Retrieval-Augmented Fact-Checking** system that could "retrieve data in almost real time, to consult with databases."

Audience Management

Audience management refers to processes supporting the publication and wide dissemination of fact-checking content. Audience managers increase engagement by employing SEO optimization, online advertising, and conversion of written content to short videos [195]. Interviews revealed that audience management involves connecting with consumers of fact-checks over many channels, of which social media is one.

Participants used generative AI to both connect with existing audiences and reach new audiences. P33 used generative AI for **Metadata Generation** to support social media content, including "summaries, SEO for article publishing, title generation." P14 used generative text-to-speech models for **Format Conversion**, taking fact-checks and converting them into audio for short videos posted to "Tiktok, Instagram, YouTube shorts," noting that AI helps achieve the right volume for disseminating content via video sharing algorithms. P31 described working on AI-based **Story Translation** of their fact-checked content into multiple languages, a goal echoed by P14, who envisioned translating their short videos.

Participants also described efforts to connect immediately with audiences, before misinformation could go viral. P32 described a beta of a fact-checking chatbot in a **Conversational Tipline** using OpenAI's GPT-4, from which they collect circulating misinformation from users and instantaneously deliver information to audiences who use more private platforms like WhatsApp, rather than Facebook and X. Other participants, including P1 and P17, described in-progress chat-based tiplines via which users can submit suspected misinformation. P13 described an in-progress tool for **Live Fact-Checking** that "can do claim matching while a person is speaking," allowing for claims to be debunked in real time.

Participants envisioned tools for automatic web content generation and predictive analysis of audience engagement. P28 envisioned automatic **Web Page Generation** that could produce fact-checks "based on social media posts that are verified [...] and then just code the iFrames for us to be able to embed it in our own content," saving programming labor. Finally, P12 envisioned a system for **Audience Analysis**, providing insight into how audiences would consume a factcheck, and recommending it be presented as a video, an infographic, or a short or long-form article.

Technology

Technology refers to work building and maintaining data science pipelines used by fact-checking organizations. While not all organizations have a Technology unit, many uses of generative AI would be invisible without specific reference to the work of software engineers and data scientists employed by fact-checkers and their partners.

Participants described in-progress generative technologies to improve the core functionality and end user experience of fact-checking data science pipelines. P2 used generative AI for **Data Preprocessing**, to "get a rewriting or a restatement of the claim that's a bit cleaner" than unprocessed social media content or tipline messages. P8 noted that generative models improve on **Core NLP Tasks** over finetuned BERT models—"the previous generation" of NLP—including for claim content matching. **Data Summarization** for human end users was another predominant use reported by participants. For example, P2 described using generative AI to provide a human-readable summary of clusters of misinforming content, describing "the variety of content" and "how it's changing over time." P25 tested the capability of Google's generative models for natively **Handling Multimodal Misinformation**, wherein the relationship between text and image must be parsed to understand subtle misinformation or hate speech. P2 described ongoing experiments for **Addressing Data Scarcity**, noting that they would "generate pseudo labeled data" where real, "gold standard" data for novel misinformation did not exist, and either "use those labels directly or use them to train a lower cost classifier." Finally, P16 envisioned a generative model for **Detecting**

Coordinated Inauthentic Behavior and influence operations, noting AI might "do more meaningful work on [detecting] coordinated networks, behavior."

Advocacy

Advocacy refers to long-term processes to influence information policy, forge connections between fact-checking organizations, and engage with the public via misinformation literacy campaigns [195]. Participants performing advocacy work were often senior managers who also managed teams of investigators.

Participants suggested generative AI could encourage information literacy, promote relationships between organizations, and improve access to information. P8 envisioned generative AI helping in **Connecting Fact-Checking Organizations**, by standardizing the methods and technologies to combat misinformation in Europe, noting that recent models handle most European languages well. P12 envisioned generative AI scaling fact-check operations by connecting organizations across Africa: "I think that's one area in which generative AI can really help. If fact-checkers are working together [...] they can help scale the impact of their fact check to different segmented audiences that they serve [...], whether it's local language, whatever format." P21 envisioned generative AI for **Promoting Information Literacy**: "people will have the option to kind of play games with the chatbot that are intended for media literacy on misinformation." Finally, P30 envisioned generative AI **Changing the Tenor of Fact-Checking**, shifting the way audiences consume fact-checking content:

> Changing the way users can consume reliable and good information could be incredibly beneficial for fact-checkers [...] If [...] they need and get good information [...] with a chatbot, for example, or with other ways, that would be fantastic. [...] people value us as we are because we give them reliable information and they know they can trust us, but if they also knew that they can consume [that] information in any way they wanted to, I think it would be an incredible leap forward.

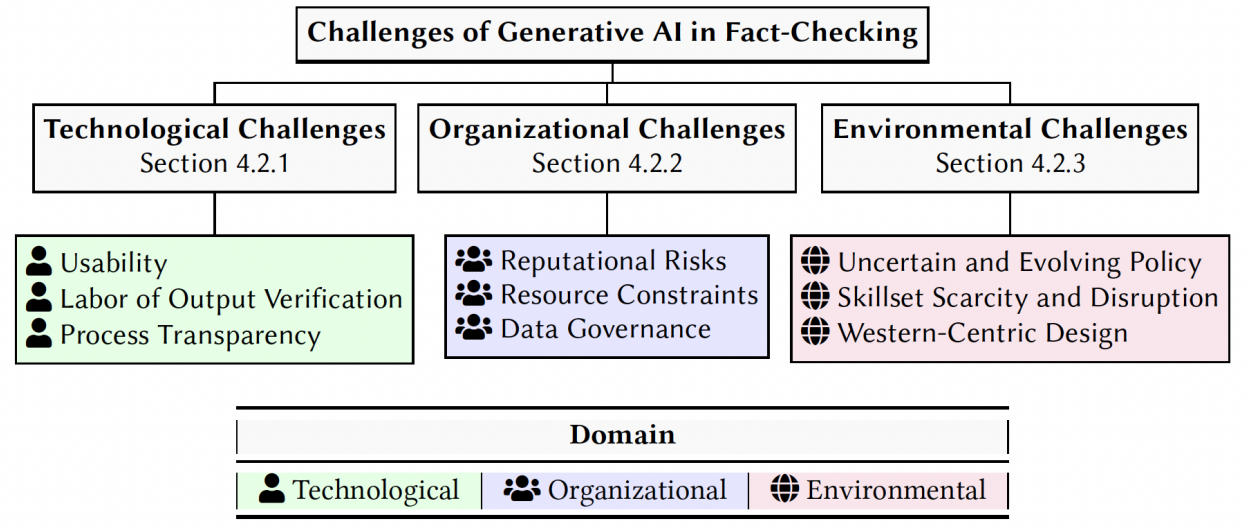


Figure 5.3: A description of the Technological, Organizational, and Environmental challenges to generative AI in fact-checking

5.4.2 RQ2: Challenges and Limitations

I found during thematic analysis that, unlike RQ1, challenges did not break down based on organizational infrastructure. Rather, participants described challenges related to using the technology itself; to incorporating it in an organization; and to factors that affected society as a whole, and were often out of their organization's control. As illustrated in Fig. 5.3, the Technology-Organization-Environment (TOE) framework [313] offers a ready model for these findings, and I use it to describe participant challenges as follows: **Technological Challenges** that impact the user of a system, such as the manual labor of verifying generative model outputs, denoted with an icon of a person (note that Technological challenges are in fact sociotechnical, involving human interaction with technology [458]); **Organizational Challenges** that impact an organization in the aggregate, such as reputational risks incurred by using systems that hallucinate, denoted with an icon of multiple people; and **Environmental Challenges** that impact not only an organization but an entire society, such as the scarcity of skilled workers, denoted with a globe icon.

Technological Challenges

Participants described barriers related to model usability; the labor of verifying model output; and the conflict in using a hard-to-explain technology in a process requiring absolute transparency.

Usability: Participants noted a lack of clarity concerning prompt engineering and hyperparameter tuning. P27 described an iterative process of choosing prompts for OpenAI models that resulted in uncertainty: "We prompted and we coded a bit and we thought, oh, this prompting technique, and combining this prompt with that one and iterating it, and then majority rule. And we [...] thought, okay, is this really the way that this should be used?" P2, who develops AI for fact-checking, noted that generative AI is "incredibly sensitive" to prompts, and "I'm sure we don't have the best approach" to prompt design. P8 noted the verbosity of ChatGPT reduced its usefulness for fact-checking content, which is "laser-focused." P14 noted that OpenAI models used for summarization randomly enter "loops" of repeating one word. P25 tweaked settings like temperature to improve reliability, but the effects were hard to see in model output. Finally, P13 expressed frustration over failures of image models like DALL-E to render text in images, limiting their use in creating visual content for stories.

Labor of Output Verification: Every participant described human review of AI-generated content as non-negotiable for ensuring the quality of published fact-checking content. Participants described some uses of generative AI as currently untenable due to the verification labor required. P11 summed up participants' opinions: "Of course it's not as accurate. The tools are not as accurate. You still need to corroborate the information that you get." P25 noted that while it is "tempting to use [generative AI] for speeding up your work," its unreliability means fact-checkers must "see [if] this is correct, what is the source?" P36 noted that, even if ChatGPT provides a lead or answers a question, "it's just as quick for us to go and find it [...] we're just so used to that lateral sort of work. And to be honest [...] we'd be going and double checking all that anyway." P14 noted that, like the internet before

it, generative AI re-organized the efforts of fact-checkers to fit the technology, noting "we had to train them" on writing styles that led to better outcomes with generative AI.

Process Transparency: Participants described generative AI as a potential impediment to the transparency needed to create trustworthy content. P29 noted, "Our sourcing is [...] always actually quite transparent. [...] we fill our story with hyperlinking to our sources and [...] how we got to everything." P8 noted that hallucinations prevented them from using ChatGPT: "The result was largely unusable [...] The sources have to be very well integrated [...] it just doesn't work. Sounds very good, [but] there will be hallucinations, it will just make up sources." P12 said explaining research that uses generative AI is hard because "as fact-checkers, we actually do not understand the processes" of the models. Finally, P35 noted that generative AI may engender questions about bias concerning selection of experts: "Have [models] weighed whether [...] there are uncertainties about them, have they been disgraced for some reason? [...] humans have biases as well, but I think in factcheck [...] there's always many, many different sets of eyes on our checks and the experts we use."

Organizational Challenges

Participants described organizational barriers including reputational risks in unpredictable technology; resource constraints preventing investment; and concerns over data provenance and ownership.

Reputational Risks: The most common organizational barrier participants identified was the reputational risk of a mistake in a fact-check. P16 said "just by default we need to be much more cautious than almost anyone else, because it's hard for us to come back from a big mistake." P8 and P26 both noted that 90% accuracy is insufficient for fact-checkers, whose relationship with audiences depends on offering information verified by experts. P2 noted that sharing generative technologies with partner organizations also shares risk, prompting further caution: "It's their organization's reputation that's at risk, not ours only." P18 said they would not trust generative models when there is only one correct answer. Finally, P2 noted that academic evaluations of generative AI are unreliable, as fact-checkers handle *novel*

information: “In an academic context, it’s always retrospective. [...] you put that into the Bing API or Google and you find lots of relevant content that can help refute that claim. But when it first appeared [...] I don’t think that was the case.”

Resource Constraints: Participants said that most fact-checking organizations lack the financial resources to invest in generative AI. P3 noted that donor funding informs the building of new technologies: “We rely on donor funding a lot, and donor funding is to address a specific use case. So if there’s not enough resources here marked for building a machine learning model, then, we just do that out of pocket or partner with other organizations. So that has been a main limitation for us.” P27 notes that, even at their well-resourced organization, “we don’t have really this AI development department” and that colleagues in low-resource organizations only develop AI tools with universities. P13 said that their organization cannot afford tools developed by better-resourced fact-checkers: “[I] met the team of <Organization> at the last IFCN conference, and they told me it’s going to be a huge sum to get that subscription [...] a small company with seven people, [we] might not be able to afford that.” P21 noted resource constraints facing organizations in the global south:

> We’re in the global south, so sometimes the resources are not the same, either to use generative AI [...] for investigating or creating our own tools. For example, some colleagues in Spain have a chatbot [...] and right now we’re trying to find resources to buy the chatbot, the basic form of the chatbot [...] it’s more like an economic problem and it’s not exclusive for [us], but probably more small fact-checking organizations in different global south countries.

Data Governance: Participants expressed concerns about the privacy of their data and the provenance of AI training data. P27 noted using open source models when possible, as they “have very sensitive material [...] investigative reporting and investigative stories, and we don’t want this to be used in models and as training material.” P27 appreciated the “legal safety” of European data protection laws, recalling a conversation with a fact-checker who highlighted the importance of trust between organizations: “I really would like to work with

all these Google tools, but still, it's Google and I'm kind of hesitating. I wish the New York Times would've developed it. Then it would be very easy for me to trust it." P31 noted many organizations questioned if they should be compensated for the labor of producing content used to train generative models: "You'll just continue to see more interest in using fact-check information to feed AI [...] Are we going to be compensated?" P12 described the uncertainty of what data was used to train models as "problematic" for fact-checkers.

🌐 *Environmental Challenges*

Participants described society-wide barriers including uncertain government and partner policy; skillset scarcity and disruption; and western-centric design.

🌐 **Uncertain and Evolving Policy**: P16 contended that, though evolving, government policies in Europe are not equipped to deal with generative AI, as well as the new forms of misinformation arising from it. P13 noted that law around generative AI in India is limited by technicalities and intended primarily to stop citizens from being "cheated" by deepfakes. P26 expressed concern about "impinging on the freedom of speech" if generative AI were overapplied for moderating speech, including that produced by AI models. P8 noted uncertainty about the policies of networks like the IFCN, saying "I think you can't just use that too much in your work, if you want to stay within the framework, which is super important for us."

🌐 **Skillset Scarcity and Disruption**: Participants consistently noted the short supply of generative AI skills. P35 said this scarcity rendered them unaffordable: "people who do know how to do that are working in organizations where they're on a much higher wage than anything that a fact-checking or journalism organization could offer." P3 said finding tech workers in Africa with generative AI skills was difficult and made harder by the headhunting practices of U.S. tech companies, noting "getting the right skill at a level we can actually afford as an NGO has been a [...] major challenge."

Many participants reported attempts to build generative AI skills internally, in part to prepare for its disruptive impact on fact-checking workflows. P25 described generative AI

skills as important both for efficiently dealing with misinformation and for awareness of the misinforming content that generative AI enables. P27 noted difficulty incorporating generative AI due to fears of displacement by the technology, and of changes to the fact-checking profession: "I think the biggest challenge is how to communicate this in a department [...] It won't replace us, probably, but it will change our work tremendously." P28 characterized generative AI as a generational issue for local fact-checkers, noting that aversion to new technology hampers the ability to match promulgators of misinformation: "We are not even yet at the point of being comfortable appearing on videos [...] if you [...] confidently face your cell phone and record what is happening, the same way malign actors are, maybe we would have a fighting chance. And I'm not even at the point of talking about AI yet."

🌐 **Western-Centric Design**: P2, P4, and P11 each said that, while generative AI tools perform well in English, their quality remains poor for low-resource languages, particularly local African languages. P13 described the local languages they dealt with as a low priority for companies developing language models. P12 highlighted the need for Africa-centered partnerships to overcome western biases and lack of access to the data for training African models:

> The challenge we have right now is, it's like the new shiny toy and everybody is talking about AI in Africa, but when you actually ask from an African point of view [...], the problem comes with [...] integrated bias in Eurocentric, or let me just say American platforms. So Meta, Google, they're all developed there and then they're used here. So the inherent biases in that, that's not something that we can do anything about [...] it'll be good to find ways, at least for African tech, to partner with these people, to develop the tools that would work for the continent. Otherwise, it's just noise.

P28 connects AI language with inauthentic colonial power: "In post-colonial places like ours, we are taught to read in another language. It's monotonous. And what I'm getting at is that that monotony is the same sound you hear in AI, and it appears very authentic, even if it's not [...] It really affects the way we are being shaped as a country."

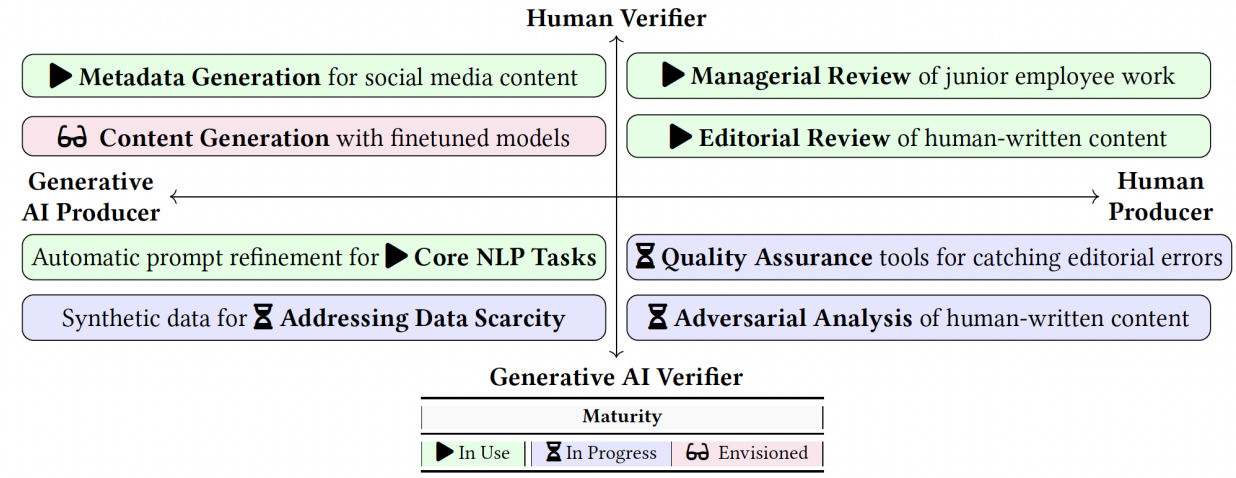


Figure 5.4: Designing for verification: A sociotechnical verification space for the production and verification of content.

5.5 Discussion

Realizing opportunities of generative AI in fact-checking requires not only building new technical competencies, but also addressing responsible use in a domain concerned primarily with reliability. I introduce a novel sociotechnical dimension to the design space for generative models that centers information verification; discuss tensions between the values of fact-checking and the values of generative AI; and outline a research agenda for generative AI in fact-checking.

5.5.1 Design Considerations for Generative AI in Fact-Checking

Designing for Verification

Every participant centered one concern about generative AI: verification. This arose in Organizational challenges, in reputational risks of publishing unverified output, and in Technological challenges, as fact-checkers must transparently explain their processes and verify model output.

To begin addressing these challenges, I introduce a sociotechnical dimension to the design space of generative models focused on the production and verification of content, conceived using a 2x2 matrix. I locate the *Producer* of content on the X axis, and the *Verifier* of content on the Y axis (see Fig. 5.4). The upper right (*Human Producer, Human Verifier*) characterizes workflows in most fact-checking and digital media companies, as experienced staff review content authored by junior staff, and editors refine content authored by fact-checkers and writers. The bottom right (*Human Producer, GenAI Verifier*) characterizes the ⧗ **Quality Assurance** system for catching small errors (P27), and the ⧗ **Adversarial Analysis** of fact-checking content (P3). It adds security for high-stakes tasks like fact-check publication that are performed primarily by humans. Fact-checkers generally agreed that the bottom left quadrant (*GenAI Producer, GenAI Verifier*), which includes no human oversight, is suitable for low-stakes settings, or where there is a clear evaluation metric that can be used by the verifying model. For example, P2 used generative AI to refine prompts used by other generative models in ▶ **Core NLP Tasks**, improving pipeline performance with little human oversight. Finally, participants held mixed views of the upper left quadrant (*GenAI Producer, Human Verifier*), noting that the 👤 **Labor of Output Verification** may render such designs inefficient, or devalue the role of the human. Participants welcomed AI for ▶ **Metadata Generation** and ▶ **Format Conversion** of content into new modalities, noting such uses saved time, even with human review.

Verification is essential for domains where generative AI may provide transformative benefits, but where the consequences of incorrect output are high. Consider applications of generative AI in law [393], where high-profile mistakes have rendered use of generative models suspect, or medicine, where research suggests demographic biases emerge in models trained on medical text [4]. The verification dimension, envisioned for the high-stakes, information-centered domain of fact-checking, provides a framework for conceptualizing applications that value veracity at least as highly as efficiency.

IFCN Principle	IFCN Description	Value	Tension with Generative AI
Non-Partisanship and Fairness	"Signatory organizations fact-check claims using the same standard for every fact-check. They do not concentrate their fact-checking on any one side. They follow the same process for every fact-check and let the evidence dictate the conclusions. Signatories do not advocate or take policy positions on the issues they fact-check."	Fairness	Generative AI exhibits wide variance depending on the structure of a user's input, and may reflect both implicit and explicit societal biases based on its training and fine-tuning data.
Standards and Transparency of Sources	"Signatories want their readers to be able to verify findings themselves. Signatories provide all sources in enough detail that readers can replicate their work, except in cases where a source's personal security could be compromised. In such cases, signatories provide as much detail as possible."	Transparency	Models cannot affirmatively identify sources, and may hallucinate inaccurate sources of information when asked to do so.
Transparency of Funding and Organization	"Signatory organizations are transparent about their funding sources. If they accept funding from other organizations, they ensure that funders have no influence over the conclusions the fact-checkers reach in their reports. Signatory organizations detail the professional background of all key figures in the organization and explain the organizational structure and legal status. Signatories clearly indicate a way for readers to communicate with them."	Accountability	Models are pretrained using poorly specified data, and may be aligned via practices exploiting low-cost workers in developing countries. Developers deny responsibility to compensate producers of web-scraped data.
Standards and Transparency of Methodology	"Signatories explain the methodology they use to select, research, write, edit, publish and correct their fact-checks. They encourage readers to send claims to fact-check and are transparent on why and how they fact-check."	Explainability	Even when models produce a correct answer, they cannot give a reliable explanation of how it was arrived at.
Open and Honest Corrections Policy	"Signatories publish their corrections policy and follow it scrupulously. They correct clearly and transparently in line with the corrections policy, seeking so far as possible to ensure that readers see the corrected version."	Openness	Generative AI opens a channel to disseminate information not easily observed or corrected by experts.

Table 5.2: I leverage the IFCN principles to identify fact-checking values, and participant insights to describe tensions with generative AI.

Describing Value Tensions

Value tensions refer to cases wherein the values of a stakeholder in a technology or process come into conflict with the values of another stakeholder or of the technology itself [138]. Drawing on Birhane et al. (2022) [44], I do not view generative AI as value-neutral by default, and I conduct a conceptual investigation into the tensions between the values of fact-checking in interaction with generative AI [247]. I use the five principles of the IFCN, an organization founded to promote common standards in fact-checking [311], as a basis for examining the values of fact-checking. Building on insights from interviews with IFCN signatories and partners, I map each IFCN principle to an underlying value, and describe its tension with generative AI. Table 5.2 describes these values and value tensions in detail, illustrating conflicts between generative AI and core fact-checking values such as fairness, transparency, and explainability. While many primarily technical tensions are resolvable, especially in the context of human-AI collaboration, tensions surrounding values like Accountability may require significant changes in the relationships between information professionals like fact-checkers and the technology companies that benefit from their labor [218].

Defining A Research Agenda

I describe directions for research identified by the participants in the interview study, grouped by the research community (fairness, accountability, or transparency) they primarily address (see Table 5.3). Participants stressed that researchers at universities and partner organizations will play an important role in advancing these research directions, but that research cannot have meaningful impact without the involvement of fact-checking organizations. P37 noted the limited effort of researchers to connect with fact-checkers, saying, "normally researchers are doing research without even talking with fact-checkers, so they don't know how the fact-checking world works." Some directions, such as developing technology for the global south, or combating information inequality, may also require researchers to forge new relationships with individuals and organizations outside of their existing networks.

Fairness	Directions that seek to mitigate both technical and societal bias and unfairness.	Interest
Technology for the Global South	Building technologies in coordination with fact-checking organizations in the global south to improve model performance and usability, especially in local languages.	P2, 3, 11, 12, 13, 28
Detecting and Mitigating Bias	Developing technical and human-centered approaches to identifying and minimizing bias in model output and human-AI collaborations.	P5, 6, 21, 33, 35
Combating Information Inequality	Developing methods to reach audiences outside of well-educated, well-resourced communities that typically consume fact-checking content.	P4, 9, 17, 29, 30, 31
Accountability	Directions to improve accountability of AI developers to users, fact-checkers to audiences.	Interest
Improving Data Standards and Safety	Developing technical and policy approaches to ensuring that fact-checker data and content is not misused when training or fine-tuning generative AI models.	P27, 29, 31, 37
Auditing for Deceptive Design	Auditing generative AI systems for deceptive design patterns that manipulate human users into placing too much faith in the veracity of their output.	P1, 12, 25, 26, 28
Improving Open Models	Developing highly usable open source and open weight generative AI models to alleviate fact-checker concerns related to the privacy and ownership of their data.	P2, 5, 14, 16, 27, 37
Transparency	Directions to equip fact-checkers with designs and approaches to maximize transparency.	Interest
Benchmark Development	Developing benchmarks that measure language model performance in settings closer to the real-world scenarios faced by fact-checkers, accounting for the novelty of misinformation.	P2, 3, 16, 37, 38
Synthetic Content Detection	Developing approachable generative AI tools for reliably detecting synthetic content, whatever the modality (such as text, image, audio, or video).	P3, 10, 13, 15, 22, 23, 24
Designing for Transparency	Developing design spaces and methodologies that center the transparent processes of fact-checking professionals and organizations.	P27, 31, 36

Table 5.3: I propose nine research directions for generative AI in fact-checking in which study participants expressed interest.

5.5.2 Limitations and Future Work

The findings are limited in that they focus solely on fact-checkers, and primarily on IFCN signatories and partners. There are many stakeholders in fact-checking, including audiences for fact-checks, technology providers, government bodies, and indirect beneficiaries of the impact of fact-checking on the information ecosystem [195]. Future work might center the interests and values of these stakeholders. Moreover, while I draw on the literature of organizational change, this work is primarily concerned with understanding the evolution of organizations undertaking the work of information verification, rather than organizations broadly. Future work might seek to generalize or contextualize findings with other organizations and sectors.

5.6 Conclusion

This study drew on interviews with N=38 participants at 29 fact-checking organizations across six continents to describe the opportunities and challenges of incorporating generative AI in sociotechnical fact-checking infrastructures. Insights from the interviews formed the basis for a novel Verification dimension in the design space for generative models for fact-checking. Meanwhile, the principles of the IFCN informed a description of the value tensions between fact-checking, which centers transparency, fairness, accountability, and reliability, and generative AI, an unpredictable and sometimes unreliable technology.

Human-centered design building on the design space of generative models [255] can enable the reliable use of even epistemically flawed models, helping users and organizations decide which situations lend themselves to adoption of a generative model, and when such models should not be used because they introduce epistemic risks that can compromise closely held values. Moreover, if used effectively in human-AI collaboration, these epistemically flawed models have the potential to *improve* the health of the information ecosystem, rather than further polluting it.

Chapter 6

NEEDS-CONSCIOUS DESIGN: A DESIGN FRAMEWORK USING NONVIOLENT COMMUNICATION TO MITIGATE EPISTEMIC RISK IN INTERPERSONAL COMMUNICATION

6.1 Preface

Though generative and general-purpose models have clear potential to expand epistemic risk when used in knowledge work, they also have the potential to impact human interactions and relationships by affecting the integrity of *interpersonal* information. In the research that follows, I draw on Nonviolent Communication, a well-established, process-oriented framework for improving the clarity of interpersonal communication, to create a design framework that prioritizes precise, effortful technology-mediated interactions between people, rather than low-effort interactions or scenarios that substitute people with technology. Though this study describes epistemic risk posed by many forms of online communication, generative AI emerged as a central concern among both populations of participants, and it inspired design concepts both intended to improve interpersonal connection online, and designs that, though well-intended, would effectively inhibit interpersonal connection. Ultimately, by centering its focus on interpersonal needs, Needs-Conscious Design intends to produce technologies that cut through the noise of technology-mediated communication and refocus human attention on the interpersonal information that actually matters. The findings of this chapter support **Thesis Statement 4**: *Design that centers attunement to human needs can mitigate epistemic risk when using generative and general-purpose AI in interpersonal communication.*

6.2 Introduction

Social disconnection has become one of the most pressing challenges of our time, significant enough that the U.S. surgeon general has warned of an "epidemic of loneliness and isolation" [270]. Technologies designed to capture user attention play a central role in discussions of these issues, as online communication platforms and in particular social media—a technology ostensibly designed to foster interpersonal relationships [390]—appear to contribute not to a sense of connection but to a reinforced sense of isolation for many users [317]. Creating technology that has a prosocial effect—and avoids amplifying existing problems—will require intentional design that centers human interpersonal needs, rather than simply capturing user attention, or offering a technological substitute for human connection.

In this research, I turn to a well-established model for communication that centers human interpersonal needs: that of Nonviolent Communication [347], or NVC, a process-oriented approach to empathetic communication positing that human conflict and feelings of isolation arise from unmet needs, and that communicating those needs without judgment can yield deeper connections between people [346]. NVC originated in Rogerian person-centered psychology [343] and now sees application in conflict mediation, much of which is overseen by the Center for Nonviolent Communication (CNVC) [136], which certifies trainers who teach the fundamentals of NVC. In the present work, I draw on NVC to create an approach to designing for interpersonal needs that I call Needs-Conscious Design (NCD). I address three research questions:

1. **RQ1**: How can the principles and constructs of NVC be used to produce a design framework that supports empathetic connection online?

2. **RQ2**: What design considerations and design concepts to support empathetic connection are supported by both NVC theory and the experiences of real users seeking empathy online?

3. **RQ3**: What are the design risks of a framework building on NVC that intends to

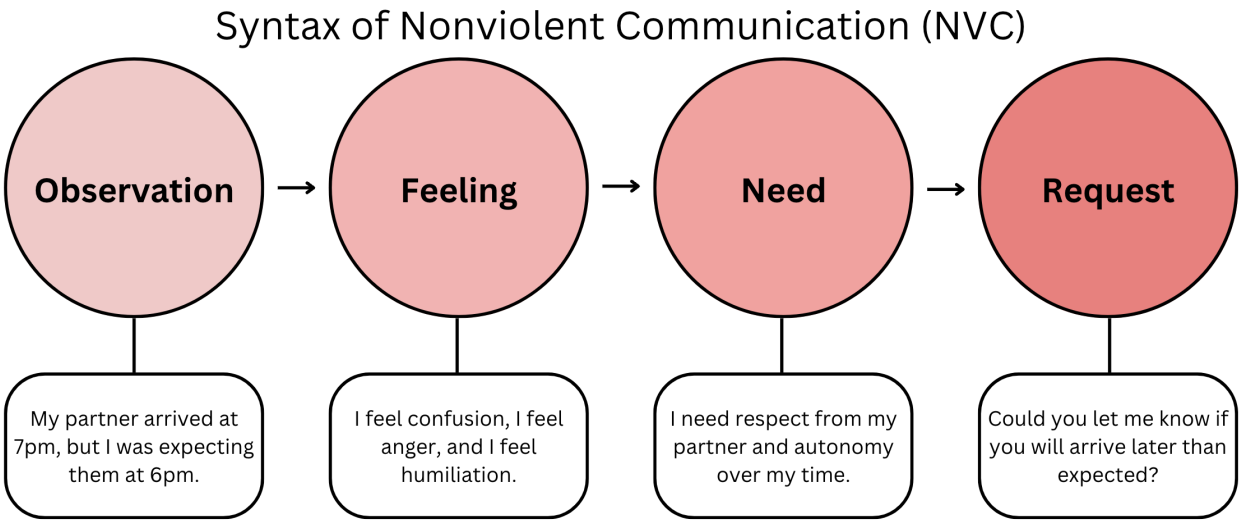


Figure 6.1: An illustration of the Observation-Feeling-Need-Request (OFNR) syntax of Nonviolent Communication (NVC), with examples drawing on interviews with certified NVC trainers.

support empathetic communication online?

To answer these questions, I enrolled N=14 trainers certified by the CNVC in an interview study. In parallel, I enrolled N=13 lay users of online communication technologies in a six-day diary study bookended with an entry interview at the beginning and an exit survey and co-design session at the end. I intended to capture both high-level perspectives from NVC trainers and the everyday experiences of users who seek empathy online. I applied a deductive-inductive coding approach [18] to the interviews with CNVC trainers and the responses and designs of lay users to derive a set of themes [53] to answer the study's research questions. I make the following contributions:

1. **I offer a conceptual model for Needs-Conscious Design, defined by the three design objectives and three levels of attunement**. Drawing on the study data, I identify supporting users in making precise observations, assuming personal

responsibility for needs, and taking intentional action as design objectives. I further define self-attunement, others-attunement, and context-attunement as progressively more complex circles in which to realize these design objectives.

2. **I provide nine design considerations for Needs-Conscious Design corresponding to the 3x3 design of the conceptual model**. Specifically, I map each of the three design objectives and each level of attunement to a concrete design consideration supported both by NVC theory and the real-world concerns of diary study participants. I further draw on co-design sessions with diary study participants to produce at least two high-fidelity design concepts illustrating how each of these design considerations could be realized.

3. **I identify five design risks for Needs-Conscious Design based on the perspectives and experiences of certified NVC trainers.** Specifically, I find that designers using NCD should be prepared to contend with challenges related to user consent, coercion and gaslighting, responsiveness to systemic harms, real-world well-being, and *empathy fog*, an emergent design risk wherein using technologies like AI to for empathetic communication obscures the effort and attention offered by a sender to a recipient.

Needs-Conscious Design offers a model for designing for empathetic connection online, grounded in a longstanding approach to meeting human interpersonal needs. Amid an increasingly polarized [215] and AI-driven [127, 121] landscape, NCD foregrounds authentic connection and mutual satisfaction of needs, providing a foundation for designs that leverage technology not as a replacement for human-to-human connection, but as a facilitator and supporter of such connection.

6.3 Methods

I conducted a diary study with N=13 participants who reported using text-based online communication at least daily, and an interview study with N=14 CNVC-certified trainers.

My University's IRB approved this research.

6.3.1 CNVC Trainer Interview Study

Interviews with CNVC-certified trainers lasted between 30 minutes and two hours, with most interviews lasting approximately one hour. Throughout the paper, I notate trainer comments with the prefix *T* (*e.g.,* "T3 said. . . ").

Interview Protocol

I created a semi-structured interview protocol that asked trainers about the following:

1. The four-component NVC model (Observation, Feeling, Need, Request).

2. Strategies and exercises for effectively using NVC.

3. What successful NVC looks like, and how CNVC trainers identify progress in their trainees.

4. Challenges in learning and effectively using NVC.

5. How NVC might inform online communication, and limitations of online settings for employing NVC.

6. What designs for more empathetic online communication might arise from NVC.

I also asked clarifying questions when participants raised other topics germane to the study (for example, I asked follow-up questions when T14 suggested that generative AI could be useful for NVC training).

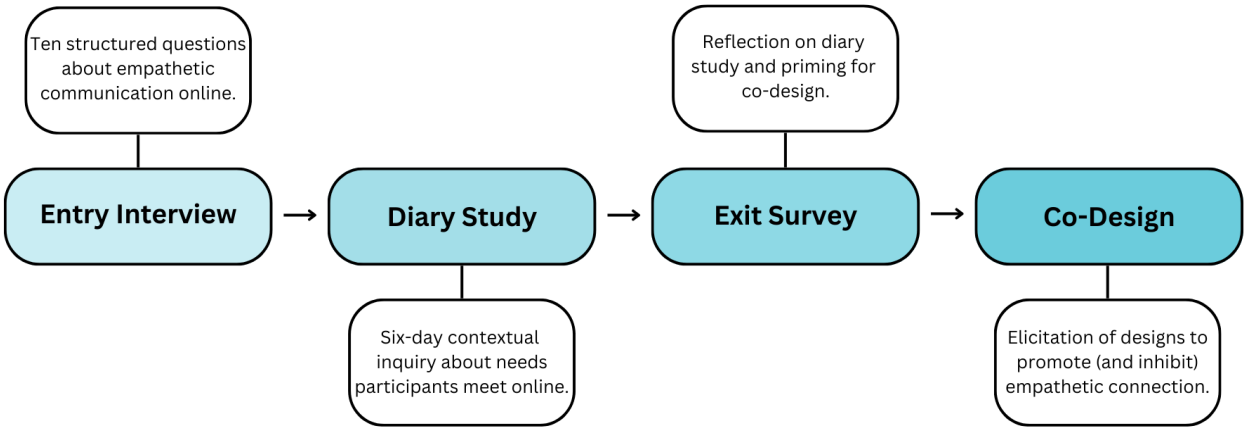


Figure 6.2: The structure of the diary study with lay participants, which began with an entry interview followed by a six-day diary study before concluding with an exit survey and co-design session.

Participants

I used the CNVC website to filter a list of certified NVC trainers in the U.S. who spoke English, in accordance with the inclusion criteria described in the IRB. I then reached out to 30 certified trainers via email, explaining my interest in interviewing them and providing an overview of the study. I provided these expert participants $50 Amazon credit and did not request that they provide demographic information.

6.3.2 Lay User Diary Study

I conducted a four-phase diary study with N=13 participants, notated with the prefix P (*e.g.*, "P3 said…"). The study began with an entry interview, then asked participants to submit diary entries for one week, and concluded with an exit survey and co-design. To prepare for the study, I piloted it first with a member of the research team and subsequently with a relative of another research team member, allowing me to correct issues with data collection forms and revise confusing interview questions.

Entry Interview

I designed a ten-question structured interview protocol for the entry interview. Four questions asked about a time participants had 1) extended empathy online; 2) received empathy online; 3) sought but not received empathy online; and 4) not extended empathy to someone who sought it online. Questions sought to elicit the interpersonal needs participants hoped to satisfy via online communication, and to surface any scenarios wherein they did not feel comfortable exchanging empathy online. The remaining questions asked about the relative difficulty of empathetic connection online vs. in person; easy vs. difficult social situations for connecting with empathy online; whether some platform designs make it easier to extend empathy; times when showing empathy is less important online; and what communication styles encourage empathy online.

Diary Study

After the entry interview, participants were asked to submit two diary entries per day via a Google Form. The first entry would describe a time during the day when the participant sought empathy online, and the second would describe a time when someone else sought empathy from the participant online. Participants summarized these interactions and shared the full text of conversations (with names and PII redacted) if they were comfortable. I gathered 109 total entries, an average of 8.4 per participant (about 1.4 per day).

Exit Survey

The exit survey asked participants about their experiences during the the diary study. Participants were asked to describe when they sought empathy and received it, and when they did not receive it; when it was easy to think in terms of needs and feelings underlying online interactions, and when it was not easy; with whom it was easiest and hardest to converse with empathy; when they chose not to communicate their needs and feelings; aspects of the online environment that made it easier or harder to communicate empathetically; and what stood

out most to them from their week of submitting the diary entries. Survey questions mirrored those of the entry interview to encourage observations that might only become salient to participants after six days of contextual inquiry [334]. A second page of the survey primed participants for the co-design by asking them to reflect on thirteen technologies envisioned by the study team, which were based on NVC teaching methods described by Rosenberg and Chopra (2015) [346].

Co-Design

The study concluded with a co-design session conducted via Zoom video call. Participants met individually with me or with the second author and reflected aloud on design features that would facilitate more empathetic communication online. After arriving at a design, the participant sketched a low-fidelity prototype using pen and paper or Zoom Whiteboarding. The participant then also reflected on features that would result in *less* empathetic communication, in order to surface designs to *avoid*. Participants then sketched a low-fidelity prototype of this design as well.

Participants

I posted study advertisements to Reddit, LinkedIn, Facebook, and to several Slack channels at my University. I enrolled 15 participants who met the study's inclusion criteria. Upon enrollment, all participants were provided with a consent form outlining the study procedures and providing detailed information about payment and study timelines.

One participant completed only the entry interview, and a second completed the entry interview and one day of the diary study. I compensated these participants for the parts of the study they completed. N=13 participants completed all phases of the study. I used the pilots to estimate participant time commitment and provided compensation according to the schedule in Table 6.1. Compensation for the co-design session was higher to motivate participants to complete the full study. Participants received Amazon gift credit, not cash, and could earn an additional $5 by completing optional fields that asked them to explain

Study Phase	Time Commitment	Compensation
Entry Interview	30-40m	$10
Diary Entries	10-20m/day	$5/day ($30)
Exit Survey	30-40m	$10
Co-Design Session	30-40m	$25

Table 6.1: The compensation schedule used for the diary study.

their perspectives in the exit survey. Table 6.2 describes the demographics of the study participants. I happened to over-sample people who identify as women and people aged 25-34.

Category	Participant Demographics (N=13)
Gender	7 Women, 3 Men, 1 Nonbinary, 1 Woman and Nonbinary, 1 Prefer Not to Say
Race	5 Asian, 3 Black, 3 White, 1 Asian and White, 1 Prefer Not to Say
Ethnicity	12 Not Hispanic or Latino, 1 Prefer Not to Say
Age Range	7 25-34, 3 35-44, 1 18-24, 1 45-54, 1 Prefer Not to Say

Table 6.2: Self-reported demographics of diary study participants.

6.3.3 Data Analysis

The study team applied a deductive-inductive approach [18] to coding the study data using the Atlas.ti qualitative analysis software [16]. Please note that, despite the AI capabilities of the software, no generative AI features were used at any point during the data analysis process. Deductive codes mirrored the research questions and included: Needs supported online (RQ1), Self-care and Self-empathy (RQ1), Patterns of empathetic communication (RQ1), Design Considerations (RQ2), Patterns of non-empathetic communication (RQ3), and Challenges in empathetic communication online (RQ3). The study team generated inductive

subcodes within these deductive codes (for example, "Needing a Space to Vent" was a subcode for the "Needs supported online" deductive code).

The second author, the last author, and I together coded all data collected from two diary study participants and met to discuss the initial set of inductive subcodes. The second author and I then independently coded data (including co-designs) submitted by one diary study participant each. We then met to propose new subcodes, merge redundant subcodes, and exchange notes and relevant quotes. We repeated this process five times to code all participant data, before meeting with the last author to discuss the final codebook. The same process was applied to code the CNVC trainer interviews, with data from two participants first coded jointly, and the remaining twelve coded by the second author and I, two participants at a time. Using the deductive and inductive codes, the we followed a thematic analysis process [54] and wrote memos on themes answering the research questions. We then met to discuss the themes, grouping them according to question, with explanatory notes and supporting quotes. To ensure that we captured both the theoretical grounding of NVC and the practical experience of individuals in the findings, we intentionally foregrounded design considerations for which we found support both in trainer interviews and in the observations of lay participants. Finally, we met and agreed on the set of final themes reflected in the Findings.

6.3.4 High-Fidelity Design Concepts

After thematic analysis, the study team used the low-fidelity prototypes and participant comments from co-design sessions to produce the higher-fidelity design concepts present in the Findings section. Note that because I conducted co-design sessions on Zoom, some user sketches were blurry or poorly realized in comparison to the full verbal description given by the participant, especially when using Zoom Whiteboarding. I include participant quotes where possible to illustrate the participant's vision for the design. The study team also identified several potential designs from NVC trainer interviews. Where trainer quotes about these designs occurred in the findings, we also created high-fidelity versions of these designs

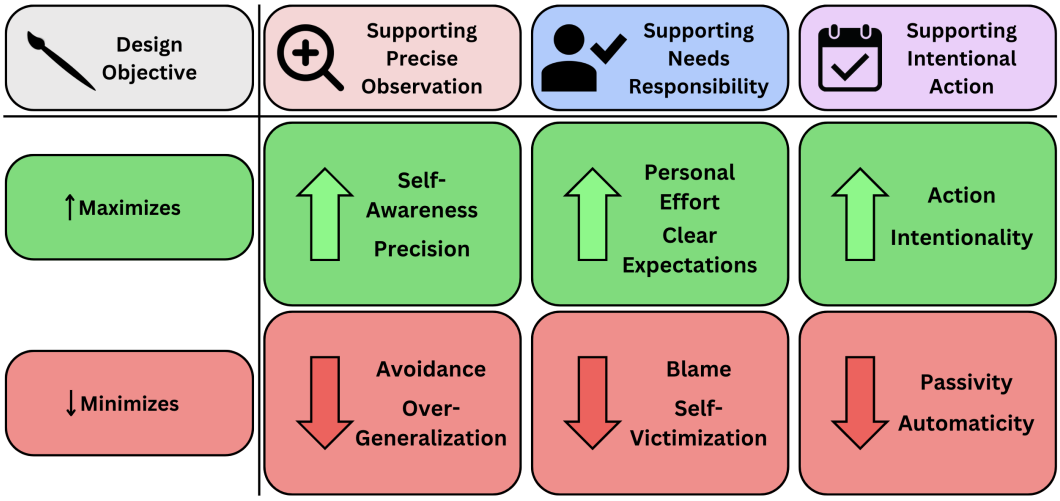


Figure 6.3: A grid illustrating the three design objectives of Needs-Conscious design. Each objective is the header of a column, and rows describe what the design objective intends to maximize and minimize.

for inclusion in the present work.

Eleven of the thirteen co-design sessions produced a positive (empathy-promoting) design, and eight produced a negative (empathy-inhibiting) design included in the present work. Note that I excluded some co-designs that ultimately did not have enough relevance for a framework intended for meeting interpersonal needs, such as an empathy-inhibiting design that exposed the user to sensationalistic news stories. I created high-fidelity versions of four designs envisioned by NVC trainers during interviews.

6.4 Findings: Conceptual Model of Needs-Conscious Design

The thematic analysis produced three primary design objectives for Needs-Conscious Design, as well as three circles of attunement in which to locate those objectives.

6.4.1 Design Objectives of Needs-Conscious Design

This research considered not only the principles of NVC that would produce better communication online, but also what would result in actionable objectives that could be translated into real designs. The data contributed by diary study participants (including their co-designs) thus proved useful in grounding the analysis in what can be realized. To that end, I identified three concrete design objectives that support meeting interpersonal needs, emerging from the studies with NVC trainers and lay participants. These objectives include:

1. **Supporting Precise Observation of Needs**: Helping users to precisely describe their experiences and name their needs, and to become more aware of and responsive to their emotional experiences and those of others. I use a magnifying glass icon to represent this objective when discussing design concepts.

2. **Supporting Responsibility for Needs**: Helping users to assume responsibility and exert the emotional effort necessary to meet their needs, including by creating spaces for connection and by finding ways to approach an interaction that support connection. I use a user icon with a checkmark to represent this objective.

3. **Supporting Intentional Action to Meet Needs**: Helping users to speak and act with authenticity and in accordance with their own intentions, including by designing for slower, less algorithmically accelerated interaction. I use a calendar icon with a checkmark to represent this objective.

Figure 6.3 further illustrates what each of these design objectives seeks to maximize, as well as what it seeks to minimize. These objectives can also be viewed as helping users to overcome common obstacles to meeting interpersonal needs online, including 1) failing to acknowledge or adequately describe one's needs; 2) locating the responsibility for one's interpersonal needs outside of oneself; and 3) failing to take appropriate action to meet one's needs after understanding what they are. As discussed in the next section, these design objectives can be approached with several circles of attunement in mind.

6.4.2 Circles of Attunement in Needs-Conscious Design

The levels of attunement defined by Needs-Conscious Design draw inspiration from NVC practice. When trainers addressed questions about how individuals make progress in learning NVC, they described several stages of NVC practice and awareness. The first was an internal practice of NVC, wherein an individual learns to recognize and acknowledge their own needs. For example, T6 identified that the first step in learning NVC as *"an internal application of Nonviolent Communication, sometimes known as self-empathy or self-connection."*. T1 highlighted the importance of being receptive to needs for achieving connection with others and with oneself: *"People who have maybe never, one, known that they had needs; two, thought that needs were okay to have; and three, ever expressed their needs, begin to acknowledge, and accept, and express [them]... that opens up a real connection in relationships."* This typically precedes practicing NVC with another person, at which point one can recognize and help to satisfy the needs of others.

Trainers also described a more advanced form of NVC practice wherein one appreciates that the syntax of NVC is merely a vehicle to what trainers describe as Nonviolent Consciousness or Needs Consciousness. This stage is characterized by the ability to adjust appropriately to context and to negotiate between competing needs. Trainers noted that embedding nonviolent consciousness into one's life can mean disregarding NVC's OFNR syntax in favor of what trainers referred to as informal NVC or colloquially as "Street Giraffe" (the giraffe is sometimes used as a symbol of NVC), wherein an individual uses less structured language to dynamically surface and respond to needs.

The three stages of NVC practice identified from the interviews inform the three circles of attunement used to structure Needs-Conscious Design. By attunement, I mean awareness and responsiveness to emotional state and underlying interpersonal needs. As shown in Figure 6.4, *Self-Attunement* is the first circle of attunement, corresponding to internal practice of NVC. *Others-Attunement* is the second circle, corresponding to practice with others. *Context-Attunement* is the last circle, corresponding to the embedding of needs consciousness

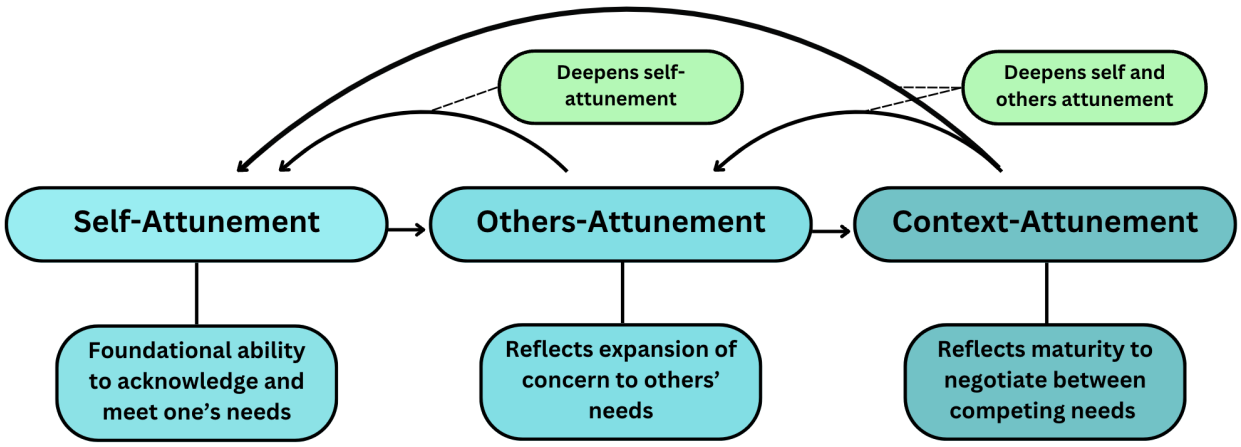


Figure 6.4: An illustration of the three levels of attunement of NCD, expanding from Self-Attunement to Others-Attunement to Context-Attunement. Backward-pointing arrows illustrate that outer levels of attunement can yield benefits for inner circles as well.

in one's life, and the maturity to negotiate between competing needs.

Both in NVC and in the model for Needs-Conscious Design, progressing to subsequent circles of attunement can yield benefits for the practice of inner circles. Learning to attune to others, for example, can yield benefits for learning to better attune to oneself. Learning to attune to context and embedding needs consciousness in one's life can yield benefits both for attuning to oneself and for attuning to others. I illustrate this in Figure 6.4 using arrows pointing from the outer circles back to the inner circles.

6.4.3 Conceptual Model of Needs-Conscious Design

By combining design objectives and circles of attunement, I form the conceptual model of Needs-Conscious Design illustrated in Figure 6.5. Each of the three design objectives of NVC is illustrated using a different base color: pink for Precise Observation, blue for Needs Responsibility, and purple for Intentional Action. The three concentric circles represented in the conceptual model correspond to the levels of attunement, with Self-Attunement as the innermost circle and Context-Attunement as the outermost circle, reflecting the expansion of

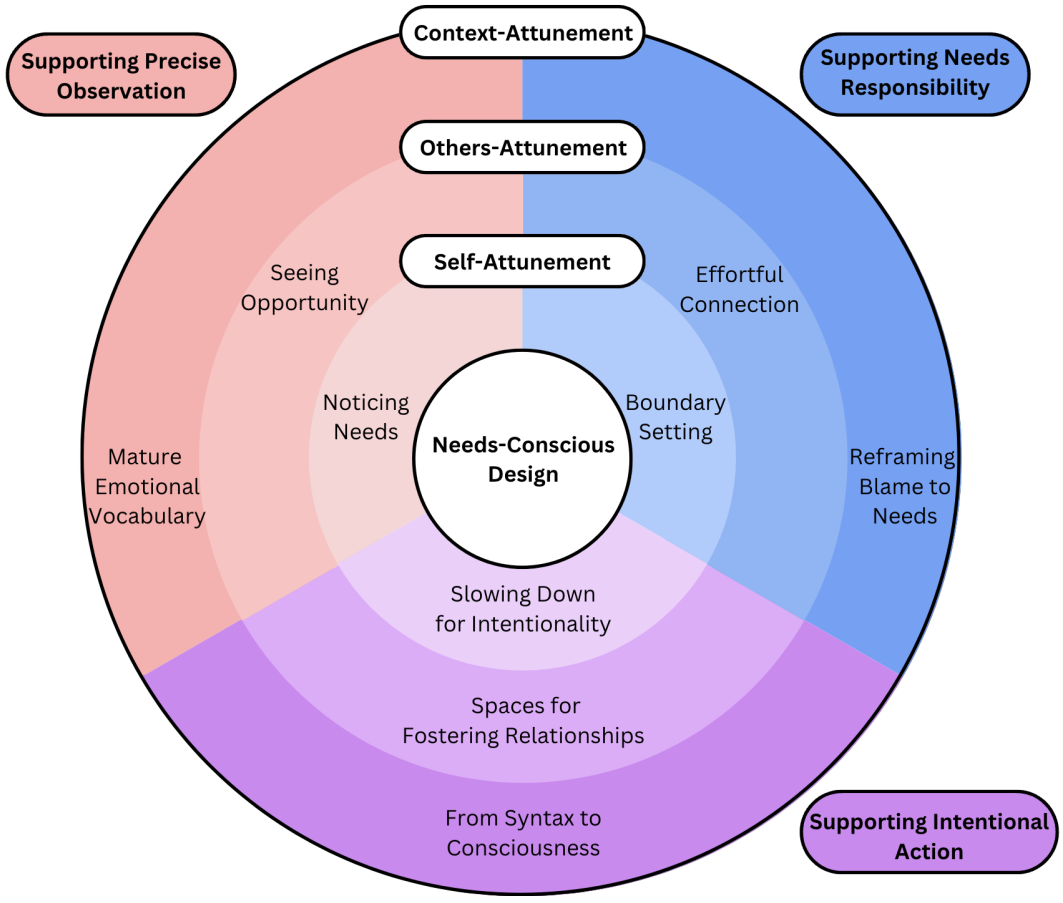


Figure 6.5: A conceptual model of Needs-Conscious Design illustrating the interaction of design objectives (mapped to pink/blue/purple hues) and circles of attunement (mapped to circles shaded progressively darker as they move away from the center of the diagram).

concern associated with each level. I also shade the outer levels a darker hue of the same color as the inner levels in the conceptual model to illustrate this progression, as well as to communicate the consistency of purpose across the levels of attunement.

Each of the circles of attunement can be accordingly divided into three arcs, each corresponding to a design objective of NCD. Within each of these arcs is written the title of a design consideration discussed in Section 6.5. I intend for the conceptual model to help designers in producing designs that help users to meet interpersonal needs, whether they are attempting to connect with themselves, or trying to better understand and communicate with others.

6.5 Findings: Design Considerations and Concepts

I discuss design considerations for NCD by first considering Self-Attunement, followed by Others-Attunement, followed by Context-Attunement, in accordance with the expansion of NVC practice. For each circle of attunement, I offer design considerations for each of the three design objectives of NCD, drawing extensively on perspectives offered by participants and presented using a descriptive title intended to summarize the consideration. I conclude each section with at least two high-fidelity design concept illustrations created based on co-design sessions and interviews with participants. Design concept illustrations are labeled with 1) the design objective supported by the concept; 2) the circle of attunement corresponding to the concept; and 3) a green plus marker indicating a positive design that supports the objective or a red minus marker indicating a negative design that undermines the objective. The name of each design concept is included below the panel in which it is illustrated.

6.5.1 Self-Attunement

Precise Observation: Supporting Needs Awareness

In the circle of Self-Attunement, designing for precise observation entails helping users recognize and acknowledge their own emotional and interpersonal needs. However, becoming

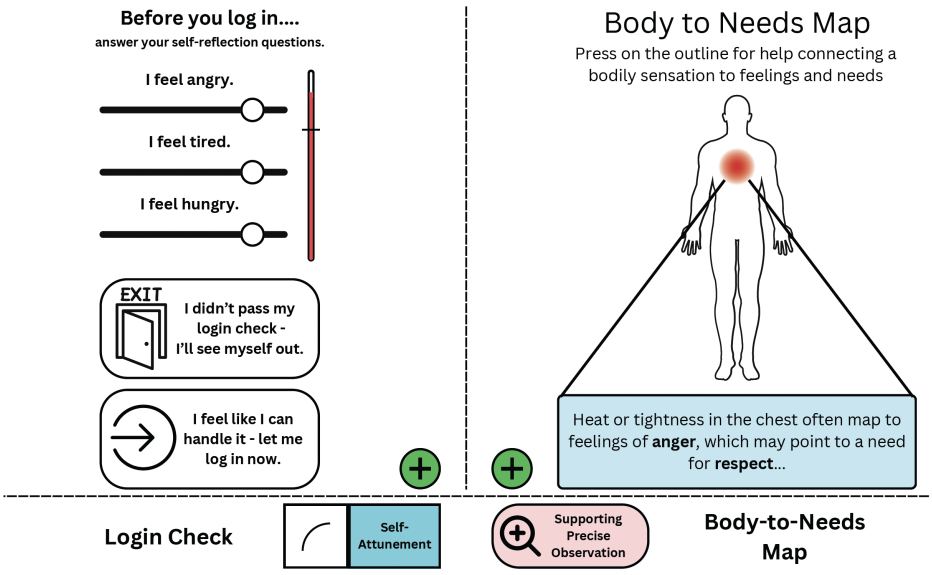


Figure 6.6: Left: Visualization of the *Login Check* design, which asks users a series of questions about their emotional state before logging in. Right: Visualization of the *Body-to-Needs Map* design, which helps users map sensations in their body to feelings and needs.

aware of what one's needs actually are can be a point of difficulty, according to NVC trainers, who said that needs can be obscured by several common false equivalences. T14 concretely identified two of these, the first of which is equating an evaluation or judgment with a feeling, resulting in a *faux feeling*: "I've asked a room full of people, who knows what it feels like to feel attacked?... One person feels angry. Another person feels scared. Another person feels confused. Another person feels disappointed when they see themselves as attacked. So attacked would be a faux feeling." T14 describes a second false equivalence caused by a failure in *"differentiating between stimulus and cause. The thing that happened outside is a stimulus, but it's not the cause of my feelings."* These false equivalencies can direct attention away from an unmet need and toward a judgment (faux feeling) or other stimulus, preventing an individual from taking action that actually meets the need.

To help become more aware of needs, trainers described exercises that link bodily sensations (such as tightness in the chest, or feelings of heat) to feelings and ultimately to unmet needs,

providing individuals with a tangible way of mapping from sensation to need. T5 said, *"in our classes, one thing we practice is being able to name the feelings and then sort of actually feel how they feel in your body physically. And of course, the most important thing is to use the feelings, stay with them just long enough to identify your need."* T8 similarly said, *"for example, sadness often can be felt in the body as an emptiness or a heaviness and an emptiness. Whereas happiness is like a lightness and a buoyancy. And anger is like a tension that involves movement. You got to move. That's why you slam doors."*

Design Concepts: Study data produced several designs intended to focus the user's awareness on their own internal state, rather than potentially algorithmically induced distractions. P1 envisioned the *Login Check* design illustrated in the left panel of Figure 6.6, which presents a user with a series of questions about their emotional state answerable using sliders, and which appears before they log into an online communication platform. P1 envisioned this as a way to become more aware of their internal state before entering an online space replete with distractions. During their co-design session, P1 described *Login Check* as follows:

> *It'd be like five questions... and you can answer them quickly... but it's also... let's take a pre-K pause... if you realize, you know what, I don't feel like doing this, then maybe you realize I don't feel like being on social media. Be kind of neat if there's a little 'leave' button.*

Drawing on interviews wherein NVC trainers like T5 described mapping bodily sensations to unfulfilled needs, the study team also created the *Body-to-Needs Map* design illustrated in the right panel of Figure 6.6. The design would allow the user to press on an area of the body and identify the sensation they were experiencing, and the interface would provide an NVC-supported explanation of the feelings and underlying needs associated with that sensation.

Needs Responsibility: Setting Expectations and Boundaries

In the circle of Self-Attunement, taking personal responsibility for one's needs entails clearly communicating what those needs are, and when those needs are not being met. Such self-advocacy has a theoretical grounding in NVC, as T1 said that NVC should help people in *"being able to speak up for what's important to them . . . self-responsibility is a huge piece. . . where's my responsibility in this experience that I'm having?. . . we begin to have our sense of our life experience is because of ourself, instead of being at the mercy of the people and events outside of ourselves."*

Diary study participants, whose observations were grounded in practical experiences, were especially aware of the importance of knowing when one's needs were not being met, and they asserted the importance of safety and reciprocity in online spaces. P3 explained that extending empathy to someone who has *"an agenda"* online can backfire and harm the person who is trying to be empathetic. P6 noted they set boundaries when requests from others involve *"unreasonable resources from the outside, like lending money, or like, coming to your house for a night."*

Participants also noted the need to set boundaries and practice self-care when one's emotional resources run low. In the exit survey, P7 expressed the difficulty of extending empathy *"when you already have emotional distress."* During the entry interview, P1 noted the importance of caring for oneself when a boundary is crossed: *"If somebody crosses a boundary, then you need to extend that kindness to yourself. You don't stop having empathy for the other person, but. . . I'm going to step away. . . It took me a long time to realize that harm none also means don't put myself in harm's way."*

Design Concepts: Co-design sessions produced two designs intended to help users set expectations and boundaries, as well as a negative design to illustrate disregard for boundaries set by a user. P5 envisioned the *Communication Style* design in the left panel of Figure 6.7 to explicitly set forth their communication preferences and expectations. Describing *Communication Style*, P5 said, *"here's the important things to know before contacting me. . . instead*

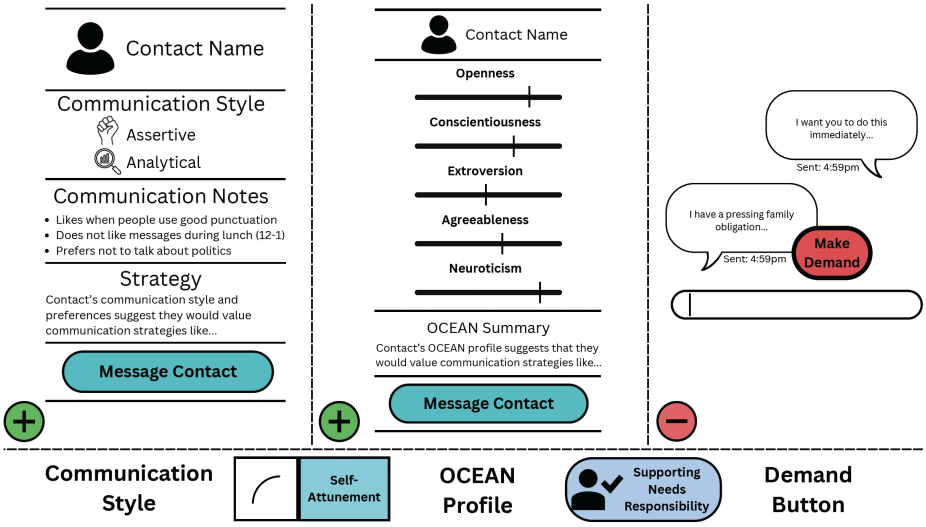


Figure 6.7: Left: The *Communication Style* design, which allows the user to make clear what to expect when communicating with them. Center: The *OCEAN Profile* design, which allows the user to provide information about their personality to other users. Right: The *Demand Button* design, a negative design that encourages a user to disregard the boundary set by another person.

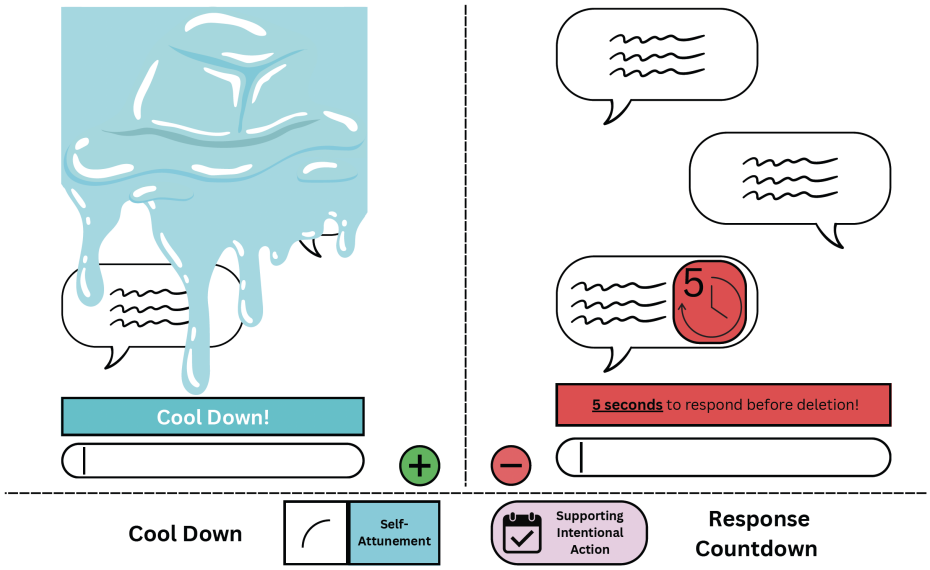


Figure 6.8: Left: The *Cool Down* design, which interrupts algorithmically accelerated communication with a calming visual element to support acting with intentionality. Right: The *Response Countdown* design, a negative design for intentionality that threatens to delete a message unless it's responded to within a certain timeframe.

of letting you just type a message . . . you to have read the first part before you actually. . . start typing." P3 had a similar purpose in mind for their *OCEAN Profile* design (center panel of Figure 6.7), which allows a user to rate themselves on the big-five traits [94, 14] so that others know how to communicate with them in a way that accords with their personality. Describing *OCEAN Profile*, P3 said, *"a person is more logical, and another person is more emotional. Talk to each other, then it's hard to connect. So it's kind of like, you've got to know the other person's language. You've got to know the other person's way." Demand Button*, the negative design shown in the right panel of Figure 6.7, also arose from the co-design with P3, who illustrated the "insensitivity" of making a significant request at work at 4:59pm. In this negative design, when the Contact sets a boundary, the user has an option to make a demand that disregards that boundary.

📅 *Intentional Action: Slowing Down to Support Intentionality*

In the circle of Self-Attunement, intentional action entails slowing oneself down to prevent avoidable conflicts and ensure intended connections are made. Trainers and diary study participants alike described the fast pace of much online communication as an impediment to intentionality. As T14 concisely explained, *"the faster the medium, the less conducive it is to connection. . . when we need to actually create an emotional connection, we're missing a ton of information."* T1 similarly noted that *"if you're gonna use text, then slow it down. Take your time, let the person know you're taking your time, that you received it, think it through."* During the co-design session, P10 reflected on the inflammatory effects of platform-induced speed in online spaces, saying, *"I think that's one of the biggest contributors to unempathetic conflict online. It's just the speed of it all. . . when I've had unempathetic conversations with family members in the past, the speed of it, and the algorithm. . . it just felt like a wrestling match with everybody in the ring."* In the entry interview, P1 described the importance of *"checking in"* with oneself when emotions run high, and becoming aware of one's own emotional state: *"I think when it's distressing, your fight-or-flight response is kicking in, and I mean, you can't run away from the thing that's in your pocket."*

On the other hand, without algorithmic incentives motivating a fast response, participants said that asynchronous communication could actually help them to give more attention to their response. P9 said, *"I can choose to react in my own way, and in my own time, and choose my words a little more carefully."* P2 similarly said that *"when your emotions are running high or you're just lashing back as a reaction, it's harder to stop that in person than with online communication. You can really take more time to step away or rethink what your words are before you send them."*

Design Concepts: Co-design sessions with diary study participants produced one positive design and one negative design concerned with the speed of communication. P10 envisioned the *Cool Down* design shown in the left panel of Figure 6.8, which they described as a way of escaping from algorithmically accelerated negative emotion: *"Some sort of calming image that pops up would be great. You know, like an ice cube... I think it would be fun if it just melted from the top to the bottom, something to kind of interrupt the pattern."* On the other hand, P12 envisioned the *Response Countdown* negative design shown in the right panel of Figure 6.8, which deletes messages if not responded to within a given time limit. Describing *Response Countdown*, P12 noted that *"I like to read things over sometimes and give them more thought when I'm trying to be more empathetic. But I don't necessarily get the amount of time... I don't use Snapchat because of this."*

6.5.2 Others Attunement

Precise Observations: Seeing Opportunities in the Ordinary

For the circle of Others-Attunement, precise observation entails being able to see opportunities to connect with others, including in everyday interactions. NVC Trainers said that simply offering one's presence facilitated empathetic connection with others, and that this opportunity for connection was almost always available. T12 said that *"empathy is really about presence... if I'm just present with what's going on inside of you... that alone is going to feel good."* T10 noted that technology is well-suited to simple, everyday demonstrations of connection and

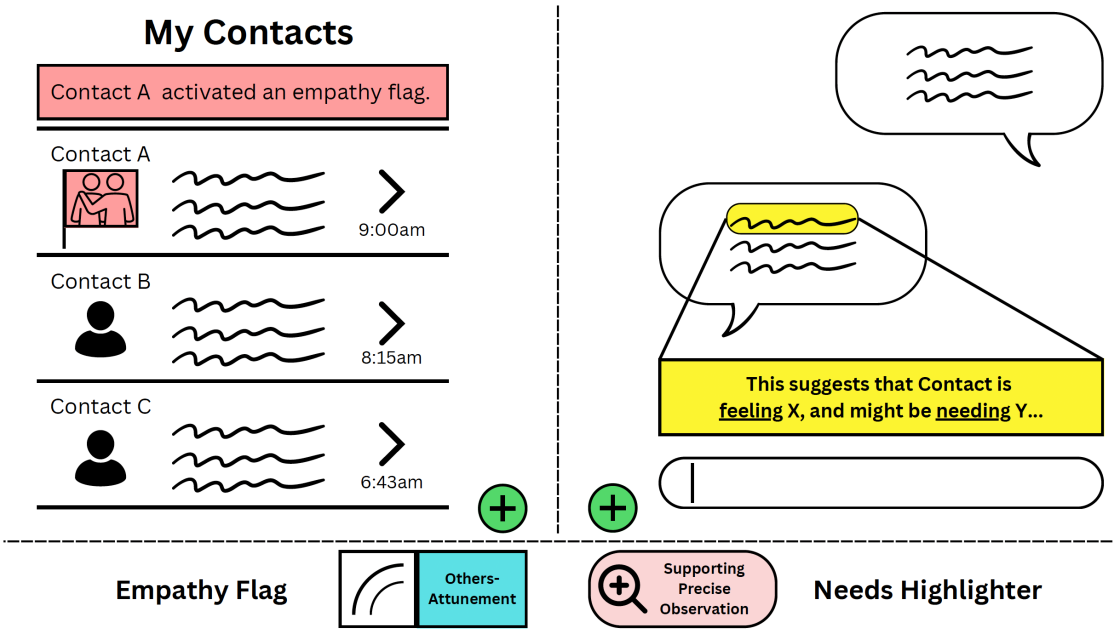


Figure 6.9: Left: The *Empathy Flag* design, which allows a user to communicate explicitly when they need empathy from another person. Right: The *Needs Highlighter* design, which supports observation of needs by helping the user to recognize them in messages.

concern: *"I think that's one of the beauties of . . . fast but thin communication. My buddy was just diagnosed with lymphatic cancer. So I check in with him and just say, hi, how are you doing? I mean, there's a lot of weight in that. That means I care about you. I'm thinking about you. I hold you dear. That's the beauty of this, technology, I think."* Similarly, when diary study participants reported receiving empathy, it sometimes occurred unexpectedly in low-stakes situations. In their exit survey, P9 described the time they felt *most* empathized with as *"the instance yesterday when I was telling my friend about how I ate too much pizza. . . really trivial but stood out because he would usually make fun of me for something like this."*

Design Concepts: Co-design sessions produced two designs focused on helping users see opportunities for everyday connection with others. P12 envisioned the *Empathy Flag* design illustrated in the left panel of Figure 6.9, which allows a user to signal when they need an empathetic connection. Describing *Empathy Flag*, P12 said that it could be useful when *"this is a serious topic, like if you're just having like a regular conversation, sometimes I'll be doing something else and just reply, and maybe not think deeply about it. But there's a way to indicate . . . something more serious."* P9 envisioned the *Needs Highlighter* design shown in the right panel of Figure 6.9, which automatically reads an incoming text and highlights sections that might indicate an interpersonal need. Describing *Needs Highlighter*, P9 said *"I drew a messaging feature where it highlights more emotionally salient pieces of the message from the other person . . . when you're writing the [response] message, I think it could highlight the components that feel the most helpful."*

Needs Responsibility: Supporting Effortful Connection

For the circle of Others-Attunement, taking responsibility for needs entails committing effort to connect with others. Both trainers and lay participants observed that effort that helps to facilitate meaningful connection. T7 expressed skepticism about low-effort connection afforded by social media platforms: *"I think about people having connection on Instagram or Facebook. . . something inside kind of makes a little checkmark, like I'm connected. I've*

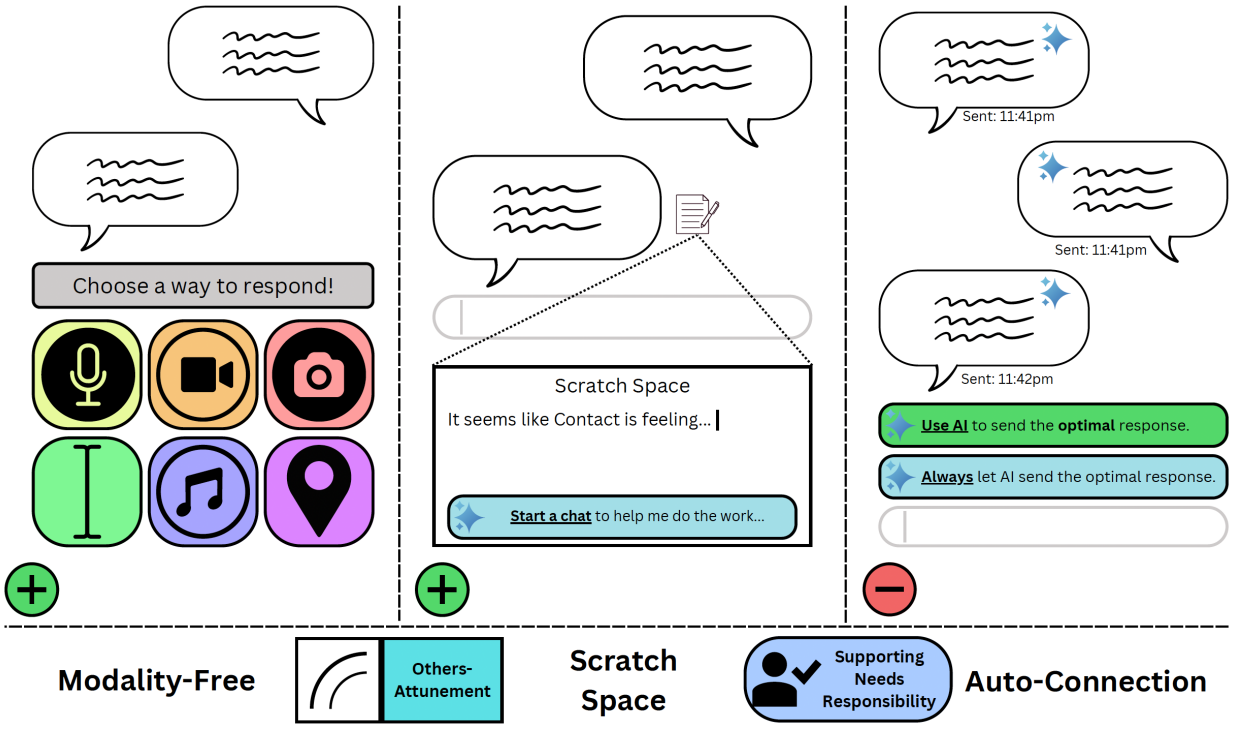


Figure 6.10: Left: The *Modality-Free* design, which encourages user effort by removing text as a default response modality. Center: The *Scratch Space* design, which encourages effort by providing an overlay in which to consider how to respond. Right: The *Auto-Connection* design, a negative design that uses AI to automate communication and marginalize emotional attunement to others.

had intimacy here. But I think the reality of it is really different... we're fooling ourselves at the nourishment we're getting." T14 noted that a fundamental part of NVC is *"to want to contribute to one another's wellbeing"*, a process that often requires significant effort to attune sufficiently to another's needs.

Diary study participants similarly noted the salutary effects of high-effort interactions. P8 described a friend's Facebook post that elicited empathy, noting that the effort the friend had put into it—indicated by its length and emotional honesty—had prompted P8 to reply with an encouraging comment. On the other hand, P12 described a time that perceived low effort created distance: *"Someone was asking for help. I told the person I always got them but they found out I was [at] an event, and the conversation went nowhere... I was unsure what happened so I felt confused... they thought I was too busy to hear them out."* P10 further described a similar effect in a different context: *"I noticed that Instagram story reactions and direct messages in response to a story can sometimes be hard to respond to with empathy, because they feel cheap."* Where effort was not clearly visible, participants often found both that they were unwilling to attune to others, and that others were unwilling to connect with them.

Design Concepts: Co-design sessions and trainer interviews produced two designs to support effort in connection, and one negative design that would undermine effort. The study team drew on a co-design envisioned by P6 to produce the *Modality Free* design shown in the left panel of Figure 6.10, which presents the user not with a default text entry bar, but with an array of different choices of modality (voice, video, image, text, music, etc.) in which to respond, encouraging users to consider how best to connect with the recipient of their message. I also drew on a co-design with P6 to create the *Scratch Space* design, illustrated in the center panel of Figure 6.10. *Scratch Space* allows the user to press on a notepad icon that fades into view a few seconds after receiving a message, and which opens an overlay for freewriting one's thoughts about the message just received. The user would also have the option of interacting with a chatbot that would ask guiding questions to help the user reflect on the conversation. Describing the ideas behind *Scratch Space*, P6 said *"you either*

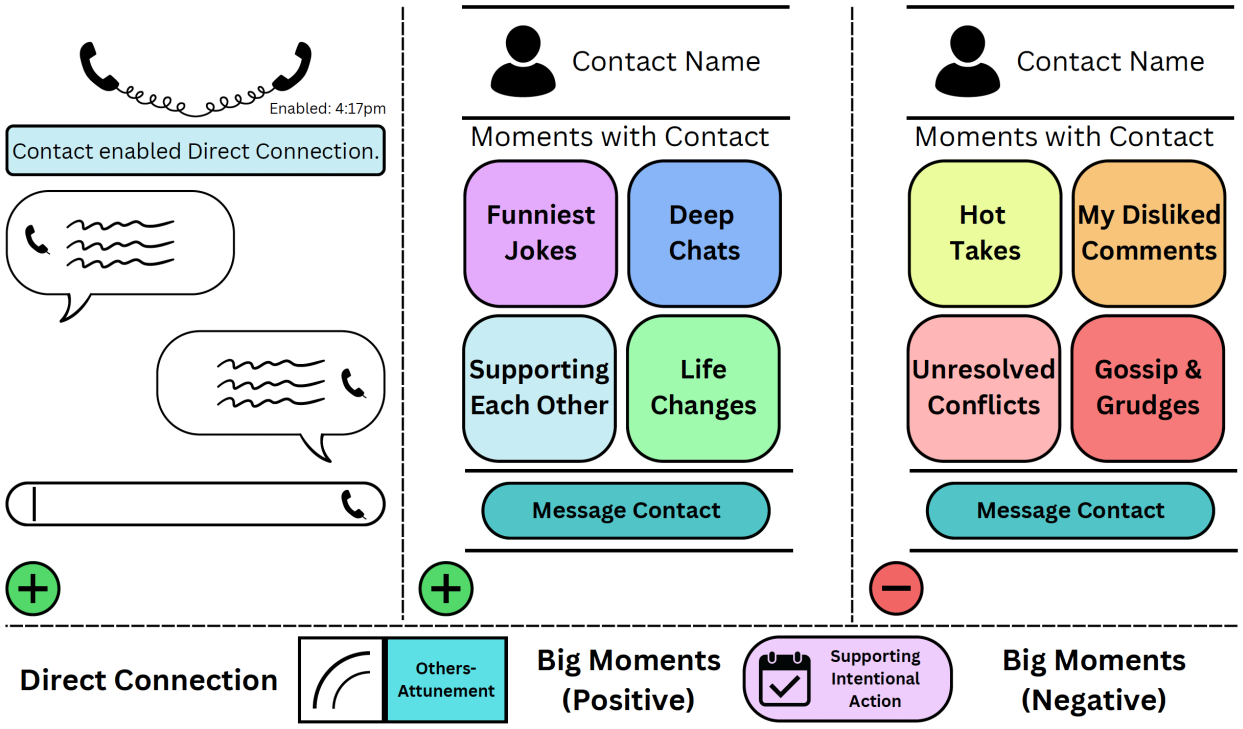


Figure 6.11: Left: The *Direct Connection* design, which creates a safe space for connection by blocking metadata collection and allowing for immediate deletion after the conversation. Center: A positive version of the *Big Moments* design, which highlights the best interactions one has had with a contact. Right: A negative version of the *Big Moments* design, highlighting unpleasant or controversial interactions.

like, freeform write down your thoughts, or have like guiding prompts to help you reflect, to kind of work through feelings and emotions. . . I think AI could be used in some cases when you just need help with having a sounding board, or just something that you can respond to."

On the other hand, P9 envisioned *Auto-Connection*, a negative, low-effort design illustrated in the right panel of Figure 6.10, wherein two users use generative AI models to automatically respond to each other. Describing *Auto-Connection*, P9 said *"it would kind of be bad if people only rely on this technology. . . so that people aren't using critical thinking to understand why they're feeling that way. . . I think the essence of empathy is understanding the deeper context."*

Intentional Action: Spaces for Fostering Relationships

In the circle of Others-Attunement, intentional action entails creating spaces in which action can be taken to build and maintain relationships. Whether with family, romantic partners, friends, or larger groups, I found that participants in the studies believed that successfully connecting in an online space requires an environment that respects the dynamics of the relationships. In the entry interview, P13 discussed maintaining connection with their family, from whom they were spatially distant, *"I don't stay over there with them, so we try to talk online. We try to communicate in whatever way we can... trying to explain our side of the story, what we're actually going through."* P9 shared a similar experience with their partner: *"my boyfriend... he's long distance... when we message each other I have to express empathy, and he also shows empathy to me."* P4 described an empathetic environment fostered by a friend in an online space during the diary study: *"during a conversation with my friend, we were discussing a personal challenge I was facing, and my friend created a safe and non-judgmental space for me to express my emotions and share my needs."* Moreover, in the entry interview, P1 said *"I see a lot of empathy exhibited in a bisexual group. It's a social community, and there'll be some thirst posts... but there's also posts from people who are going through it with their families, and they'll make it clear what they need. Like, hey guys, I really just need to talk to someone about this and probably some of you have gone through this."*

Design Concepts: Co-design sessions produced two designs concerned with creating spaces for taking intentional action in relationships. The study team drew on the co-design of P13 to produce *Direct Connection*, a fully private space for two people to connect, illustrated in the left panel of Figure 6.11. P13 noted concern with connection online because of the possibility that their intimate conversation with another person might be stored or recorded for use outside of its initial context, and that information from the conversation could be "tapped in[to]" by someone else. I drew on P13's design, which used traditional corded phones to symbolize safety from digital monitoring, to create a design concept that allows the user

to lock out any monitoring by other apps, and protects the conversation against metadata collection, also providing the user with an easy mechanism to erase the protected part of the conversation permanently after it ends.

The co-design session with P11 produced both positive and negative versions of a *Big Moments* design, illustrated in the center and right panels of Figure 6.11, respectively. *Big Moments* algorithmically curates a user's interactions with another person. The positive version of the design highlights pleasant or meaningful experiences online with the contact, while the negative version highlights experiences that will draw the user's attention, but will likely not lead to better connection. Describing the inspiration for the negative version of *Big Moments*, P11 said that *"I know some people on Reddit... they like to go through people's histories and they'll be like, didn't you say this thing... like your most disliked comment or maybe your most controversial comment."* Describing the fundamental congruency of the design for a positive version, P11 said that the *"only difference is, I think, it's my most liked comment."*

6.5.3 Context Attunement

Precise Observation: Mature Emotional Vocabulary

At the level of Context-Attunement, precise observations means using a well-developed emotional vocabulary that's appropriate to the situation. Many NVC trainers highlighted that gaining maturity in NVC practice meant developing an expressive vocabulary for needs and feelings. T14 emphasized the need for a *"mature emotional vocabulary"*, noting *"I can use my words to block you out and not let you know what's going on inside me, or I can use my words to give you a sense of what my experience is."* Referring to an inventory of descriptive words sometimes used in NVC practice, T12 encouraged trainees to *"memorize the needs and feelings vocabulary. I think that's really important."* T7 described learning NVC as analogous to learning a new language, and noted they use an NVC-inspired deck of cards [157] containing the *"inventory of feelings and an inventory of needs"* with trainees.

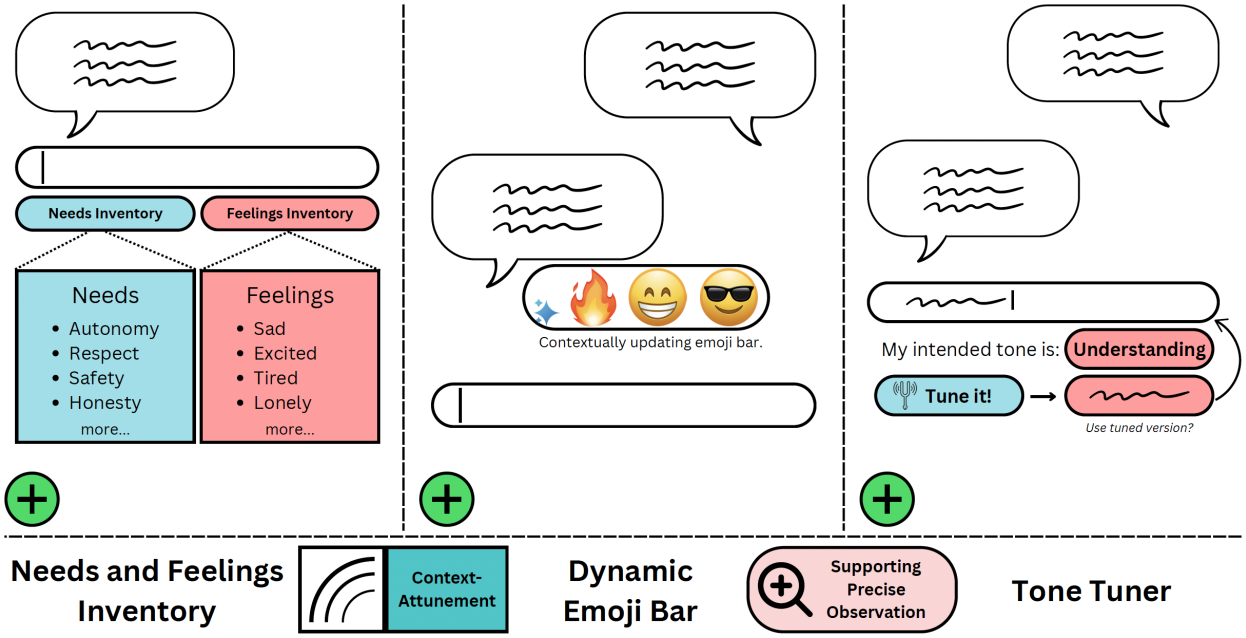


Figure 6.12: Left: The *Needs and Feelings Inventory* design, which supports the user in precisely describing feelings and needs for a given context. Center: The *Dynamic Emoji Bar* design, which uses AI to produce emojis appropriate to the context. Right: The *Tone Tuner* design, which allows the user to request AI assistance with getting the tone right for a message.

Diary study participants similarly noted the importance of an expressive vocabulary for connection online. P3 noted in their exit survey that *"with online communication, I feel like I don't need to beat around the bush... as long as there are words that indicate emotions ('this is so exciting!' , 'I love it so much!', 'thank you!') then the reader will not view my messages as threats."* P8 similarly affirmed in their exit survey the importance of a good vocabulary for expressing needs: *"Language and expression is quite important. Sometimes emoticons can be used to convey what words don't."* Conversely, participants noted difficulty connecting with others who do not clearly communicate feelings and needs. P9 said, "it's harder when people don't respond... it's also hard when even if I follow up, they don't clarify any more."

Design Concepts: The study data produced three designs concerned with employing an expressive and contextually appropriate emotional vocabulary. The study team created the *Needs and Feelings Inventory* design illustrated in the left panel of Figure 6.12 based on the interviews with T12 and T7 described above. The design includes buttons that users can press to access an inventory of needs or feelings, which can also be linked to more complete, NVC-supported definitions of these needs and feelings. The co-design with P4 produced the *Dynamic Emoji Bar* design, illustrated in the center panel of Figure 6.12, which uses generative AI to dynamically create emojis appropriate to the context. P4 echoed P8's suggestion that emojis could play a role in online platforms, where words were sometimes not sufficiently expressive. The co-design session with P2 produced the *Tone Tuner* design, illustrated in the right panel of 6.12. *Tone Tuner* allows a user to select the tone they were seeking to achieve after writing out a response, and have a generative AI model suggest revisions that accord with that tone. Note that *Tone Tuner* does not function without the user first writing a response (*i.e.,* it does not write the response for the user), lest the design undermine the previously described goal of supporting users in effortful connection.

Needs Responsibility: Reframing from Blame to Needs

For the circle of Context-Attunement, taking personal responsibility for needs entails reframing expressions of blame and judgment into expressions of unmet interpersonal needs. T4 described

how NVC attempts to facilitate this reframing: *"we've been conditioned to focus on who's right and who's wrong... and we shift the question [in NVC]... rather than who was wrong, who needs what?"* T10 noted that NVC *"isn't about making them wrong and bad... it's more... the way you're meeting your needs is not meeting mine."* T8 described a strategy common in NVC *"to translate judgments to needs. The need is the opposite of the judgment most of the time, especially if it's a negative judgment... for example, I felt betrayed... I can translate that to... I have a need for trust. Trust would be the opposite of the idea of betrayal."* T2 noted that failing to make this translation, and instead assigning blame, comes at a price: *"what's the cost of blame?... blame brings defensiveness and disconnection."* Diary study participants also said that evaluations of who is right and wrong tend to inhibit empathetic connection. P11 noted in the entry interview, *"you tend to be less empathetic to others because you have like this baseline of, well, I think this is right and this is wrong."*

Design Concepts: The study data produced one design to support reframing blame to needs and one negative design that encourages the assignment of blame. The study team envisioned a *Judgment Translator* design based on the perspective of T8 noted above, as illustrated in the left panel of Figure 6.13. *Judgment Translator* alerts the user when a message they've written might contain a judgment that masks a need. It then prompts the user to write what they're feeling and what they're needing before translating that judgment into an expression of what they need. The co-design session with P5 produced the negative design *Timed Grievance*, illustrated in the right panel of Figure 6.13. *Timed Grievance* reminds the user at one-hour intervals that their most recent message has been left at seen, subtly encouraging the user to assign blame for the lack of response. Describing *Timed Grievance*, P5 said that *"you might've completely forgotten about it, and then it just brings it to the forefront."* They noted that this could be an especially negative design *"if you've said something very vulnerable."*

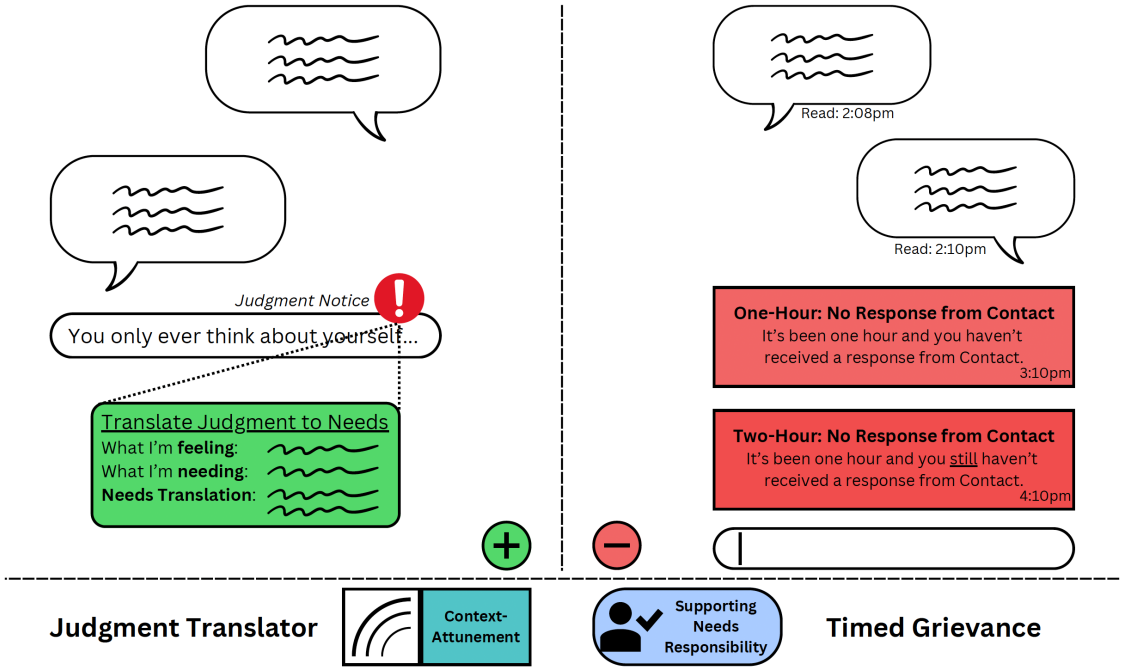


Figure 6.13: Left: The *Judgment Translator* design, which notifies the user if their message contains a judgment and helps them reframe it as an expression of needs. Right: The *Timed Grievance* design, which encourages blame by reminding a user every hour that their message was not responded to.

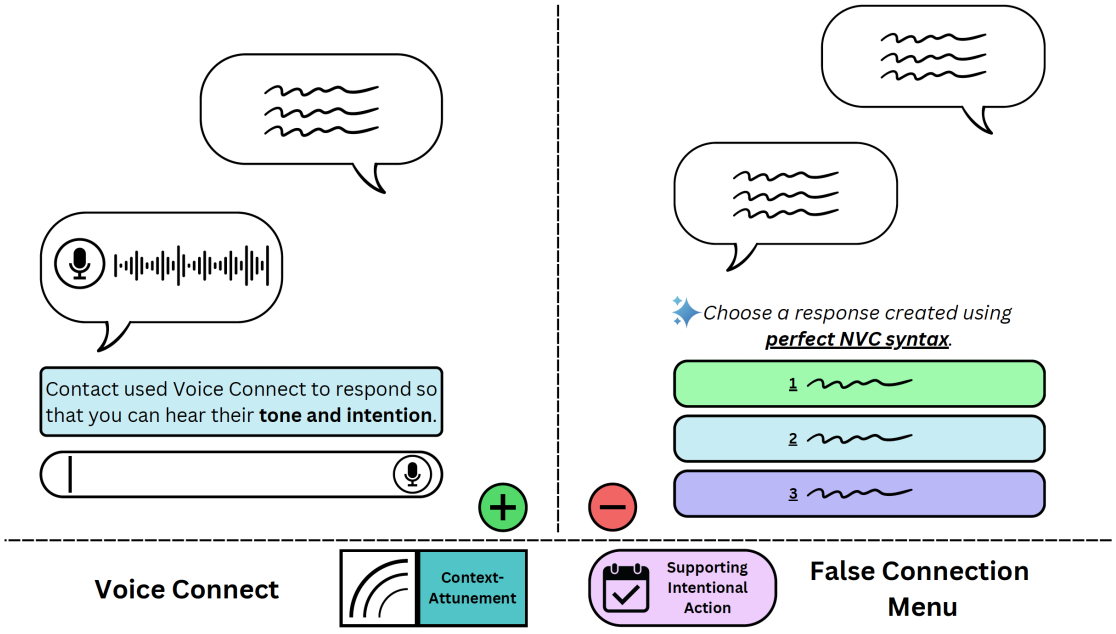


Figure 6.14: Left: The *Voice Connect* design, which allows the user to highlight that they are communicating vocally so that their tone can convey the authenticity of the message. Right: The *False Connection Menu* design, which produces perfectly formed NVC messages but divorces the message from the consciousness of the user.

Intentional Action: From Syntax to Consciousness

In the circle of Context-Attunement, intentional action entails embedding needs consciousness into the choice architecture with which one navigates everyday life. Every trainer with whom I spoke highlighted the difference between following the structured format of NVC (*i.e.*, the Observations, Feelings, Needs, Requests [OFNR] formula) and practicing NVC with authentic Nonviolent Consciousness, sometimes referred to as Needs Consciousness. T12 explained that the structure of NVC is just *"a framework that helps you get to... NVC consciousness, which is focusing on, 'I care more about my connection with you, and also my [connection] with myself."* Transcending the syntax of NVC means living authentically within the principles of NVC. T7 tried to capture this by explaining the *"core value of NVC response: authenticity, empathy, responsibility, shared power, and choice... not to get my agenda over on you, get your buy-in, or coerce you into my agenda."*

Design Concepts: The co-design sessions with diary study participants produced one design intended to help transcend syntax in favor of authenticity, and a negative design intended that highlights the drawbacks of syntax without consciousness. P7 envisioned the *Voice Connect* design illustrated in the left panel of Figure 6.14, which provides a user with the ability to send a voice message, along with a message highlighting that the message is being sent specifically so that the contact can hear the intention in their tone. Describing the intention behind *Voice Connect*, P7 said they prioritized *"voice recording because in texts I feel like people can read in different tones, and maybe misjudge the message you are trying to send."* Conversely, P2 envisioned the *False Connection Menu* negative design, illustrated in the right panel of Figure 6.14. *False Empathy Menu* automatically presents the user with three perfectly formed empathetic responses to a conversation partner. While the responses are well-formed on the surface, they are created independently from the user's consciousness, and without the user's effort. Describing *False Connection Menu*, P2 made clear the mechanical intention behind the design, noting that *"you're just presented with a list of options, kind of similar to those phone call automation flows where you're just stuck in this infinite loop*

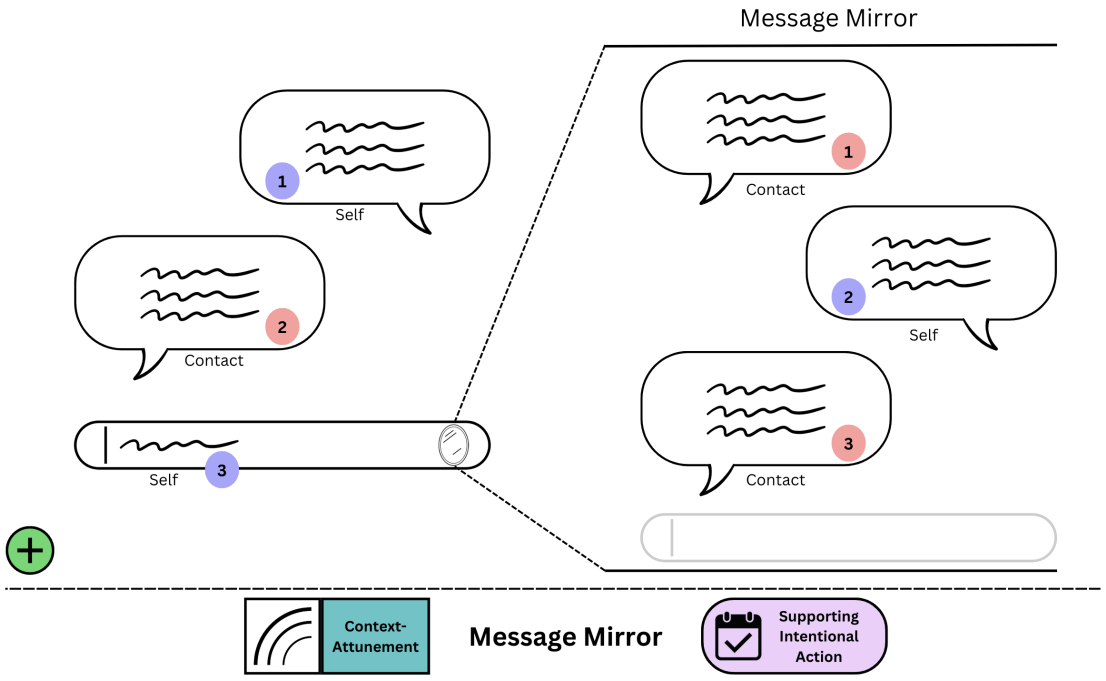


Figure 6.15: A visualization of the *Message Mirror* design, which prompts self-reflection by presenting a user with their own messages as though they had been sent by a conversation partner.

of trying to get help, but just being bounced around from place to place." As noted in the illustration, these menu options could be presented in perfect NVC syntax, but still lack needs consciousness.

Finally, the study team produced the *Message Mirror* design illustrated in Figure 6.15 based on the interview with T2. T2 envisioned using AI to help a user reason about whether one's words were intended for connection and appropriate to the user's immediate context: *"train an AI with that . . . what would it be for you to receive what you're about to say? Just that mirror." Message Mirror* includes a small mirror icon in the text input bar. When the icon is pressed, an overlay presents the user's messages as though they are the Contact's, and vice versa. The unsent message being typed by the user is presented as though it had already been sent.

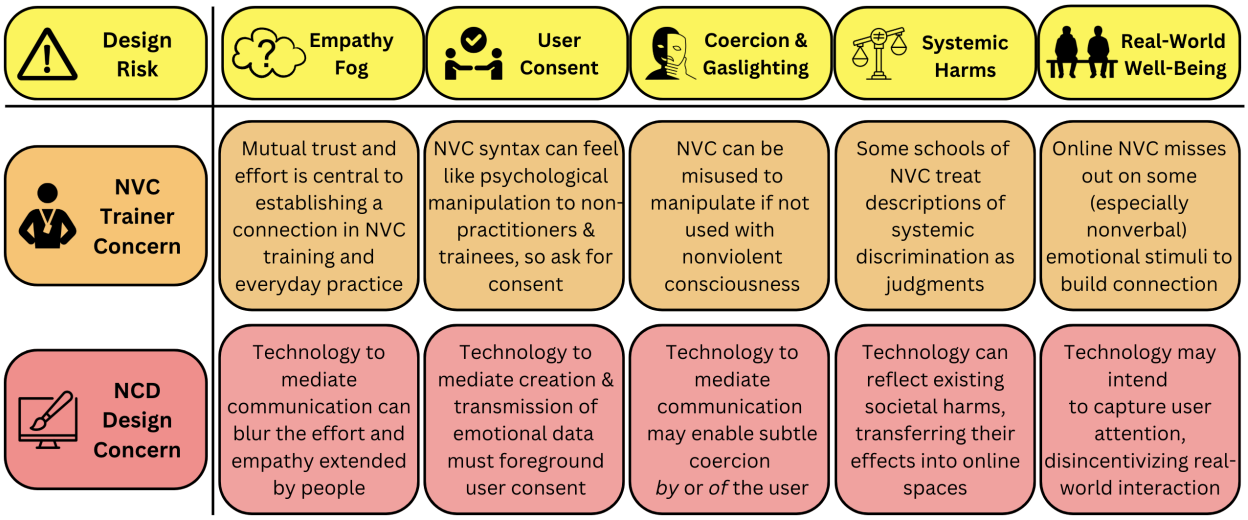


Figure 6.16: A grid illustration of five design risks of NCD identified using interviews with NVC trainers. The header row includes the design risks. The first row describes trainers' words of caution about potential misuses and mistakes surrounding NVC practice. The second row describes how these concerns might translate into design risks associated with NCD.

6.6 Design Risks

Drawing on interviews with NVC trainers, I surfaced five design risks resulting from the NCD design framework's foundation in NVC. While NVC trainers were uniformly enthusiastic about NVC, they were also careful to warn about the potential for misguided and flawed approaches to NVC, including in the context of technology. I thus discuss five design risks by describing both the concerns expressed by NVC trainers and their potential implications for NCD. Figure 6.16 illustrates the five design risks, along with the concerns expressed by trainers and the corresponding concern for NCD.

6.6.1 Empathy Fog

The first design risk surfaced by the study was a phenomenon I call *Empathy Fog*, referring to the uncertainty of whether or not empathy was in fact being exchanged in a technological

environment increasingly mediated by artificial intelligence. While the study did not directly ask participants about AI (except to seek clarification), this concern was more common and more salient than any other expressed by the trainers, who discussed the challenges to empathetic connection online posed by ChatGPT [281] and other generative AI technologies [324]. T4 expressed this concern succinctly: *"as we enter the age of AI, there's a part of me that doubts it's even you... it could affect empathy and connection in a potentially dangerous way."*

While there are risks associated with using AI to mediate one's relationship and communication with oneself, trainers were more supportive of improving interpersonal connection by using AI to improve the relationship with oneself. T7 captured this perspective when discussing the role AI-driven technologies could play in helping people to slow down and reflect: *"AI is going to be really good at that, not injecting the make-wrong thing, helping people to re-regulate their nervous system. And while they're doing that, during the pause, they're giving themselves empathy. So they're connecting to their needs and their feelings and hopefully doing the same for the other side."*

I illustrate *Empathy Fog* in the left panel of Figure 6.17, which visualizes the opacity of the connection offered when AI is used to directly mediate communication. The right panel of Figure 6.17 illustrates the more optimistic perspective offered by T7: that by helping the user improve their relationship with themselves, a design can ultimately improve their connection with others as well.

6.6.2 User Consent

The second design risk surfaced by the study was the need for *User Consent*. Trainers noted that NVC can come off as an unwelcome psychological strategy to people who don't practice NVC themselves. This is particularly true when using the highly structured OFNR syntax of traditional NVC, but it can also be the case when people are asked about feelings or needs that they're not prepared to explore. Trainers discussed obtaining consent as an important part of NVC, especially when practicing it with people who are not themselves versed in

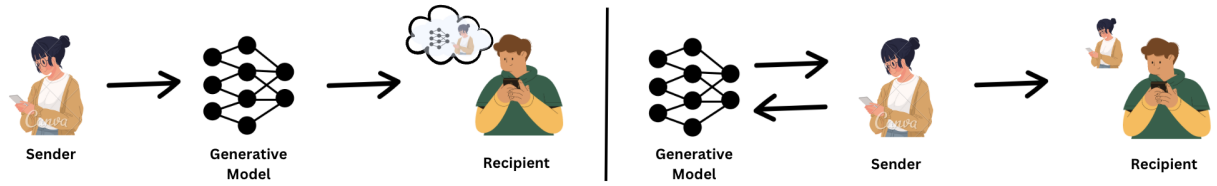


Figure 6.17: Left: An illustration of *Empathy Fog*, an emergent design risk for communication significantly mediated by technology (and especially AI) that obscures the empathy and connection received by a user. Right: A potential solution to empathy fog suggested by NVC trainers, wherein technology helps the user with self-reflection but does not directly mediate communication.

NVC. Consider the perspective of T7, who said they always asks for consent before practicing NVC with new trainees:

> "*Always ask permission... give that person permission, if you get this feeling that I'm doing this thing to you, just please let me know and we can stop right away. So, real permission, giving consent... that builds trust... honesty and transparency, in my experience with NVC, and other places and other realms, is worth the while. It's worth slowing down for.*"

Designs produced using NCD must also foreground user consent, especially given that such designs would likely involve the creation and transmission of user emotional data [93]. I suggest that the affirmative consent framework of Im et al. (2021) [182], which asserts that consent given online should be voluntary, informed, specific, revertible, and unburdensome, can be leveraged to pose guiding questions with regard to the consentful creation of designs using NCD. Even when technology mediates only between a user and themselves, the user should feel they have consented to an interaction that may inform how much of themselves they are willing to offer, including emotional data.

6.6.3 Coercion and Gaslighting

The third design risk surfaced by the study is the potential for misuse that supports coercion and gaslighting. The Consent section above intends to address concerns about potentially well-intended but nonetheless non-consensual applications of NCD. However, trainers also said that without nonviolent consciousness, the NVC syntax can be abused, potentially leading to gaslighting or coercion. T14 said that *"if you miss the consciousness part, then my intention could be to get my way, or my intention could be to manipulate a certain outcome. And then I'm using words that sound like NVC, but now I've weaponized NVC and it's not NVC at all."* Similarly, T3 said, *"NVC can be a tool that can contribute to harm if not used with the mindset of seeking connection... it's very important to use NVC with a lot of intentionality."*

Like NVC, designs produced using NCD must be mindful of the possibility that they may either enable the user to gaslight or coerce another person. As noted previously, T14 describes true NVC connection as desiring "to contribute to one another's wellbeing" and then acting accordingly. Designs that enable coercion either *of* the user or *by* the user fail to achieve the goals of NCD, in that they do not support the user in truly taking responsibility for their own needs, or in sufficiently attuning to the needs of others.

6.6.4 Accounting for Systemic Harms

The fourth design risk surfaced by the study is the potential to overlook or ignore systemic injustice. Some trainers pointed out that NVC does not necessarily enable people to better respond to systemic harms. T2 said *"there's not a complete agreement between all NVC practitioners around... how to look at social justice... some people who are like, yeah, the question of privilege has nothing to do with NVC. That's not what it's about. It's just about feelings and needs. There's a lot of other people saying, absolutely not. Privilege is essential to have a deeper sense of empathy and understanding power dynamics."* T3 pointed specifically to a debate within the NVC community about whether observing a power dynamic should be regarded as an observation, and thus a precursor to a discussion on feelings and ultimately

needs, or as a judgment or evaluation—which comes with blame attached. T3 said *"I think NVC needs to be very much informed about privilege, systemic racism... and I do believe that to go and say, well, these are not observations, is actually a harmful move. So somebody who would teach NVC that way, to me contributes to the very systems that are oppressive and harmful."*

As is clear from much prior work on algorithmic reflection and magnification of societal harms [28, 65, 328], online spaces are not immune from systemic injustice [29, 400]. The perspective of T3 above suggests that designs produced using NCD should also seek to be aware of how they contribute to addressing or potentially exacerbating existing harms in online spaces.

6.6.5 Real-World Well-Being

The final design risk surfaced by the study is the failure to prioritize real-world well-being. Despite the enthusiasm of many trainers for better online spaces for connection, they also noted that nothing could replace the richness of in-person interaction. Remarking on what's missing in online communication, T14 said, *"so much of communication is non-verbal. It's body language, it's tone of voice, it's other pieces."* While T12 noted that they embraced practicing NVC in online spaces and were optimistic about developing new apps to help people use NVC, they also said that *"having the same interaction online does not stimulate the same... there's a reason why people in the NVC community are drawn to in-person. There's an energetic exchange that you can't reproduce 100% online."*

Unlike many designs that seek to capture user attention (*i.e.*, user engagement) [235, 234, 356], NCD should not seek to maximize time on an app, but to maximize connection to others and to oneself. In many cases, that goal is better met by in-person connection and time spent away from devices than it can be by online communication. Designs for NCD should thus avoid user engagement for the sake of engagement, and help users to make the decisions that are best for meeting their interpersonal needs.

6.7 Discussion

The NCD framework builds on the tradition laid down by prior work seeking to give users more control over their online relationships and their emotional lives. However, by focusing specifically on enabling users to meet their interpersonal needs, and by drawing on the rich theoretical tradition of Nonviolent Communication and the lived experiences of diary study participants, the NCD framework differs from prior work in a few key ways.

The first difference lies in the less-is-more philosophy that runs through both design objectives like supporting needs responsibility and through design risks like empathy fog. NCD acknowledges that even though their intentions are good, designers can easily do too much when attempting to support empathetic connection online. This can occur when a design leads to decreased effort on the part of a user, or, in the case of empathy fog, to confusion about how much effort has been committed to an interaction. Technologies like generative AI constitute a double-edged sword in this respect; while they can enable designs like *Needs Highlighter* that promote attention to another person's needs, or enable more expressive interactions between individuals via designs like the *Dynamic Emoji Bar*, they also risk automating the most important parts of an empathetic interaction. Successfully designing to support empathetic connection thus requires close attention to the effects of powerful new tools on the process by which communication is produced.

A second significant difference from prior work lies in NCD's utility at several levels of interpersonal attunement, starting at the level of the individual. This mirrors the progression described by NVC trainers, wherein a trainee first practices an internal version of NVC, then applies it in interpersonal settings, and finally learns to embed the values of NVC in their life. While the levels of NCD reflect this progression, they also provide designers with an opportunity to reflect on what sort of empathetic interaction they intend to design for—does it intend to help the user to connect more with themselves, with others, or to more successfully navigate a complex interpersonal environment online? As the design concepts demonstrate, the design objectives of NCD can manifest quite differently depending on

the intended level of attunement. For example, for the precise observation of needs design objective, I offered the *Login Check* design concept at the level of self-attunement, the *Empathy Flag* concept at the level of others-attunement, and the *Tone Tuner* concept at the level of context-attunement. While more expansive forms of attunement may benefit more individual attunement, I also noted that designers need to be careful of increasing complexity and potentially unintended consequences in Others-attunement and Context-attunement, such as the potential for supporting coercion or gaslighting in the hands of a user whose goal is not to connect but to get their way.

Finally, the NCD framework differs from prior work in its emphasis on the user's personal responsibility for meeting their interpersonal needs. This is something that good design can support, including by providing spaces that give the user a space to reason through their needs in a conversation, as in the *Scratch Space* design, and by helping users to communicate personal boundaries and expectations more clearly, as in the *Communication Style* design. While an approach that foregrounds personal responsibility can help users to see the agency they have to meet their needs, it may also comes with some of drawbacks that were highlighted during trainer interviews. As noted in the design risks, the emphasis on personal responsibility can also potentially result in a failure to properly recognize and contend with systemic harms. Some NVC trainers have adapted their approach by allowing the observation step of the OFNR syntax to describe oppressive systems, and I suggest that responsible design with NCD should be similarly responsive to unjust treatment in society.

6.7.1 Limitations and Future Work

This work has several limitations. First, while I have read the authoritative background literature on NVC and have sought to represent the perspectives of certified NVC trainers accurately, I am not an NVC trainer myself, and thus there may yet be aspects of NVC not fully captured in this work. Second, while I developed NCD to be broadly applicable, some of the more concerning recent applications of technologies intended for empathetic connection online apply to specific settings. Examples of these include AI-assisted therapy

[162, 272] or romantic relationships with AI companions [217, 165], applications that might be construed as meeting interpersonal needs through self-reflection, but which carry very different implications and expectations from person-to-person communication online. How to design for settings like these could be more specifically covered in future work. Third, I relied primarily on lay participants for co-designs, drawing on trainer design ideas only when trainers explicitly suggested a design during interviews. Future work might conduct a more complete and intentional co-design study with CNVC trainers or other experts on interpersonal communication. Fourth, I note that NVC has developed into a framework focused primarily on conflict mediation and interpersonal connection, rather than explanatory theories of human behavior. While psychological theories like BPNT [406] and SDT [405] are likely compatible with NVC, I did not seek to develop that connection in the present work. Future work might seek to more concretely make this connection, or to interview therapists to produce design frameworks that connect contemporary therapeutic approaches with the methods of NVC. Finally, note that participants were all English speakers. Future work might also study the application of Needs-Conscious Design in diverse cultural contexts, as NVC is itself practiced around the world [66, 197].

6.8 Conclusion

This study introduced Needs-Conscious Design, an approach that draws on Nonviolent Communication and centers precisely observing interpersonal needs, taking personal responsibility for needs, and acting with intentionality. It showed that human-centered design can facilitate information integrity not only in factual contexts (like knowledge work), but in interpersonal contexts, where users' trust in *each other* can depend on the way in which human communication is mediated. Needs-Conscious Design offers an approach that values human-to-human connection and the satisfaction of needs, providing a foundation for designs that leverage technology not as a replacement for human-to-human connection, but as a facilitator and supporter of such connection. Such designs can help to reduce epistemic risk by prompting users to think more critically about their communication, rather than increasing epistemic

risk by automating communication and reducing human intentionality.

Chapter 7

TOWARD LABORATORY-SCALE AI: CONTENDING WITH THE HEIGHTENED EPISTEMIC RISKS OF PROPRIETARY MODELS

7.1 Preface

One of the most noteworthy risks of general-purpose AI, surfaced both in my study with fact-checkers and in my study with CNVC trainers, is that of data privacy and provenance. Users and organizations alike express concern that their data may be captured by a chat interface or an API and used for purposes they did not intend. The broader question of data and model transparency—how data is captured, manipulated, and postprocessed—also plagues the scientific study of proprietary models, for which the reproducibility of results is uncertain [275]. To make matters worse, the hardware resources needed to run open alternatives to proprietary models are both scarce and costly. In this study, I investigated whether users and organizations can extricate themselves from the heightened epistemic risk of proprietary models by leveraging small open models running on low-cost, widely available hardware. The findings of this chapter support **Thesis Statement 5**: *The heightened epistemic risks posed by closed-weight proprietary models can be mitigated by fine-tuning small open models to perform specific tasks while maintaining their general-purpose chat interface.*

7.2 Introduction

According to a report by The Verge, OpenAI's ChatGPT boasts more than one hundred million weekly users, two million developers using the API, and more than 80% adoption of among Fortune 500 companies, making it one of fastest growing services in history [309]. Despite the influence of OpenAI's flagship language models on the world's ways of working and seeking information, scientists know little about them: details of the architecture, parameter

counts, and training data of GPT-3.5-Turbo and GPT-4-Turbo are omitted or glancingly described in the company's technical reports [280]. Reaffirming the "values encoded in machine learning research" described by Birhane et al. (2022) [44], transparency has taken a back seat to values that preserve corporate competitive advantage. For many scientists and public interest practitioners, this lack of transparency is at best concerning, and often a reason to avoid such models in their work altogether [224, 293]. At the same time, recent research has enabled the use of accessible and inexpensive hardware to train domain-adapted models. Eight-bit and four-bit quantization allow very large models to run on affordable commercial-grade GPUs [106, 108]. Quantized low-rank adaptation (qLoRA) [181, 107] allow large models to be customized to a domain by adding and tuning a modest number of parameters while allowing most pretrained weights to remain fixed.

These technologies could collectively help enable a future for AI that is not wed to the interests of Big Tech corporations—one that prioritizes transparency, cost-efficiency, and the domain-specific and responsible application of language technologies, in addition to strong performance. In this work, I intend to provide an empirical, practical foundation for this approach, which I call "Laboratory-Scale AI." Concretely, I address the following research questions:

- **RQ1: Do open models offer domain-specific performance competitive with closed models for tasks of scientific and public interest?** I assess open models against closed models on three tasks selected for their scientific or public interest value: government records entity resolution [143], climate misinformation fact-checking [116], and clinical dialogue summarization [27]. I evaluate OpenAI's GPT-3.5-Turbo and GPT-4-Turbo against three open instruction-tuned models: Mistral-7b-Instruct-v.01 [191], Falcon-7b-Instruct [7], and LLaMA-2-Chat-7b [398]. Results show GPT-4-Turbo exceeds the performance of the four other models when using them in a few-shot setting, but GPT-3.5-Turbo and **open models are comparable to GPT-4-Turbo or exceed its performance after fine-tuning for a single dataset epoch**. On fact-checking,

fine-tuned Mistral-7b-Instruct achieves accuracy of .75, exceeding the mark of .72 by three-shot GPT-4-Turbo.

- **RQ2: Are open models cost-competitive with closed models?** I find that the cost of running inference on a test dataset with GPT-4-Turbo is comparable to both fine-tuning and inference using an open model. Cost savings achieved by using an open model after fine-tuning are especially notable: on the climate fact-checking test dataset, **inference is almost ten times less costly using an open model** (Mistral-7B-Instruct, $.31) than zero-shot GPT-4-Turbo ($2.65).

- **RQ3: How responsive are small open models to domain-specific fine-tuning data?** I evaluate the performance of LLaMA-2-Chat-7B fine-tuned for clinical dialogue summarization after 0%, 20%, 40%, 60%, 80% and 100% of task training data, and evaluate the performance of the LLaMA-2-Chat-7B, Falcon-7B-Instruct, and Mistral-7B-Instruct every 500 steps of the 4,298-sample fact-checking dataset. After 20% of the fine-tuning dataset (240 samples), the summarization model achieves accuracy of .79 on the test dataset, within .02 of its best. After 2,000 fact-checking samples, Mistral-7B-Instruct achieves accuracy of .71, comparable to fine-tuned GPT-3.5-Turbo. The results show **open models can be adapted with a small amount of data**, without the need for large-scale data collection.

- **RQ4: Can fine-tuned, domain-specific models provide a general-purpose chat-based interface for end users?** Chat-based language models provide an approachable interface to end users. I investigate whether fine-tuning inhibits the utility of this interface by comparing the performance of fine-tuned LLaMA-2-Chat-7B models with the base model on tasks not reflected in the model's fine-tuning dataset. For example, I measure performance of the fine-tuned fact-checking model on the entity resolution task. I find that **fine-tuned open models exhibit performance comparable to general-purpose base chat models**, and in some cases exceed it:

for example, the fact-checking LLaMA-2 improves to .85 accuracy on entity resolution, from the base model's .77.

- **RQ5: Can laboratory-scale language models be used in a responsible manner?** I evaluate open and closed models on three tasks of importance for the ethical application of instruction-tuned language models: question answering under differentially private fine-tuning, demographic bias in toxic comment classification, and abstention from answering questions for which insufficient information is available to answer correctly. I find that the performance of open models fine-tuned using a private optimizer approaches non-private fine-tuning, suggesting a better privacy alternative to closed models; that open models exhibit moderate bias that fine-tuning largely fails to mitigate; and that fine-tuning open models can improve their abstention properties: fine-tuned LLaMA-2-7B-Chat achieves an abstention score of .99 (maximum 1.0), exceeding the performance of fine-tuned GPT-3.5-Turbo. **While open models exhibit greater bias, they offer greater privacy affordances than closed models, and in some cases abstain more reliably after fine-tuning.**

The experiments demonstrate that a fine-tuned open model running on inexpensive hardware can exceed the performance of GPT-4-Turbo at lower cost. In addition to core empirical contributions, I offer a practical discussion of the challenges and opportunities of adopting a laboratory-scale approach in the Discussion section.

7.3 Approach

I review the models studied, evaluations employed, and consistent cloud environment used across the experiments.

7.3.1 Models

The models studied share the following characteristics:

- **Causal (Generative) Pretraining Objective**: All models share the causal language modeling (next-word prediction) objective introduced to the transformer architecture by Radford et al. (2018) [323].

- **Instruction-Following**: All models undergo supervised fine-tuning to enable a user to issue instructions in natural language, and receive a natural language response from the model [290].

- **7-Billion Parameters (Open Models)**: The open models each have approximately seven billion trainable parameters, allowing them to be deployed on identical cloud instances. OpenAI has not disclosed parameter counts for GPT-3.5-Turbo and GPT-4-Turbo, but studies suggest they are much larger than open models [300].

I study only generative, instruction-following models for three reasons. First, this accords with the architecture and training regimen of the closed, industry-dominant OpenAI models against which I assess open models. Second, both the closed and open models studied are among the most widely used language models in the world as of this writing, with Meta's LLaMA-2 model and Mistral's Instruct model routinely among the most popular models in the HuggingFace Transformers Python library. Third, these models provide an approachable natural language interface for users who may not be skilled in machine learning but would nonetheless benefit from the use of a domain-aligned language model. In addition to being aligned with the goal of empowering scientists and public interest users, the importance of an accessible interface is borne out by the success of ChatGPT, which far exceeds the userbase of OpenAI's own GPT-3 base models [26]. Finally, studying one group of similar models permits use of consistent infrastructure, allowing me to evaluate cost.

Defining Closed vs. Open Models

I define a *closed* model as a model which is accessible only via a call to an API, and the weights and architecture of the model cannot be accessed. An *open* model is one for which the

pretrained weights and architecture are made available and can be modified and built upon. These models are not necessarily licensed to permit any use of the model, as such licenses may still prohibit commercialization or use for unethical purposes as defined by the organization releasing the weights [398, 399]; that is, *open* models are not necessarily fully *open source* models. This definition of "open" aligns with that employed by Palmer et al. (2024) [293] and Rogers et al. (2023) [342], but omits the requirement that researchers know the data on which the open model was trained, as even in previous definitions, data requirements come with the caveat that such data need not actually be "available for direct inspection" [293].

Closed Models

I study two closed OpenAI models: GPT-3.5-Turbo and GPT-4-Turbo.

- **OpenAI GPT-3.5-Turbo**: OpenAI's cost-efficient and broadly performant model optimized to follow instructions [281, 288]. A GPT-3.5-Turbo fine-tuned with RLHF and Proximal Policy Optimization is the model available to non-paying users who access ChatGPT through the online interface rather than the OpenAI API [281]. I used OpenAI's default GPT-3.5-Turbo at the time of the experiments, which points to "gpt-3.5-turbo-0613" [288].

- **OpenAI GPT-4-Turbo**: OpenAI's state-of-the-art language model, available at greater cost than GPT-3.5-Turbo [288, 289]. GPT-4-Turbo holds the zero-shot state-of-the-art on numerous NLP tasks as of this writing, and achieves first place in human evaluations of chat-based models in Chatbot Arena [481]. GPT-4-Turbo handles much longer text input sequences (128,000 tokens) than GPT-3.5-Turbo, as well as multiple input modalities, such as images [288].

Open Models

I study the following three open models:

- **TII Falcon-7B-Instruct**: A generative model pretrained on 1.5 trillion tokens of the RefinedWeb dataset [301], released under the Apache 2.0 license by the UAE's Technology Innovation Institute (TII) in April 2023 [7]. TII's RefinedWeb dataset consists of filtered web data, and a subset is publicly available [301].

- **Meta LLaMA-2-7B-Chat**: A generative model pretrained on two trillion tokens of publicly available datasets and made available under the LLaMA 2 Community License by Meta AI in July 2023 [399]. The Chat model was fine-tuned for dialogue and underwent RLHF to improve helpfulness and minimize toxic output [399].

- **Mistral AI Mistral-7B-Instruct-v0.1**: A generative model released under the Apache 2.0 license by Mistral AI in September 2023 [191]. Mistral-7B-Instruct-v0.1 is trained on an undisclosed quantity of data from the open internet, and exceeds LLaMA-2-7B-Chat and LLaMA-2-13B-Chat on common benchmarks [191].

7.3.2 Model Evaluation

I evaluate models in zero-shot, few-shot, and fine-tuned settings.

- **Zero-Shot**: The model is provided with a bare instruction of the task, and given the data to perform the task.

- **Few-Shot**: The model is provided with an instruction and examples of how to respond. I use multi-turn formatting to provide few-shot examples to the model, following the HuggingFace chat template documentation for open models, and OpenAI's documentation for closed models. Falcon-7b-Instruct is not fine-tuned with a defined chat template, and I adhere to the guidance of the model's developers, including examples in a single user prompt.

- **Fine-Tuned**: The model is fine-tuned on a task-specific dataset before evaluation on the task's test dataset. For consistency, I fine-tune for a single dataset epoch, reporting total

examples in train and test datasets. I was unable to fine-tune GPT-4-Turbo at the time of the study, for which fine-tuning was available only via an experimental program.

Hyperparameters

I employ four-bit quantization [107] in both inference and fine-tuning. I use qLoRA adapters [181, 107] to fine-tune on domain-specific data, adopting the optimal hyperparameters specified by Dettmers et al. (2023) [107]. Specifically, I use qLoRA to tune linear layers, set LoRA matrix rank to 32, and set LoRA dropout to .05, which improves performance in models with fewer than 13-billion parameters [107]. I used gradient checkpointing during fine-tuning to save memory by recomputing activations during the model's backward pass [81, 64]. I set batch size to 1 due to memory limitations. I use the default hyperparameters for training GPT-3.5-Turbo, with the exception of fine-tuning for only one dataset epoch, rather than the OpenAI default of three.

Representative Task	Train Samples	Val Samples	Test Samples	Eval Metrics
Entity Resolution	700	100	200	Accuracy, F1 Score
Climate Fact Checking	4,298	1,842	1,535	Accuracy, Weighted F1
Clinical Dialogue Summarization	1,201	100	400	BLEU [295], BERTScore F1 [473]

Table 7.1: Representative tasks with total training, validation, and test samples, as well as evaluation metrics.

7.3.3 Cloud Infrastructure

I use a consistent cloud environment, allowing comparison of the cost and runtime of open vs. closed models. I defined the environment such that a 7-billion parameter model could be fine-tuned using qLoRA in 4-bit precision with a 1,024-token context window. I chose this setup because 7-billion parameter models are the lowest entry point for the three families of models I study (LLaMA-2-Chat-7B, Falcon-7B-Instruct, and Mistral-7B-Instruct), because

fine-tuning in four-bit precision is competitive with fine-tuning in higher precision [107], and because the tasks (*e.g.*, summarization), benefit from a context window of at least 1,000 tokens. Fine-tuning used a $0.32 per hour Google Cloud Platform (GCP) [46] spot instance with the following characteristics: a 16GB Nvidia T4 GPU; 60GB RAM; a 16vCPU, 8-core processor; and 200GB disk. While cost may vary based on region and provider, I found price was generally consistent on GCP and other providers such as AWS and Lambda Labs, within about $.05 per hour. Because I expect that most laboratory-scale AI applications will be fault-tolerant during fine-tuning, I use spot instances, which may be terminated to support higher paying workloads, but are less costly than on-demand resources.

7.4 Multifaceted Evaluation of Open vs. Closed Models

I select a practical, representative sample of tasks, including those that 1) reflect real-world uses of generative instruction-tuned models (*e.g.*, fact-checking chatbots, like Aos Fatos' FatimaGPT [131] or Meedan's Check [241]); and 2) reflected consequential work envisioned by other research. For example, Gilardi et al. (2023) [147] suggest ChatGPT can be used for data annotation (I consider specifically entity resolution), and Waisberg et al. (2023) [413] explore GPT-4 for triaging patients via clinical dialogues. I acknowledge it may not be *desirable* to use an LM in a setting like clinical dialogue summarization or fact-checking, especially without human supervision, and that the tasks are *proxies* to real-world applications.

7.4.1 Representative General Tasks

I study three tasks to compare performance of open vs. closed models, with sample and evaluation metrics in Table 7.1.

1. **Entity Resolution**: I use a custom dataset of public records to evaluate performance on Entity Resolution [143]. Given two pairs of names and addresses, the model determines whether the pairs refer to the same person. One set is derived from home deeds in Mecklenburg County, NC; the other comes from voter records. The dataset contains 1,000

Model	Scenario	Entity-Resolution		Fact-Checking		Med-Summarization	
		Acc	F1	Acc	F1	BLEU	BERT-F1
GPT-4-Turbo	Zero-Shot	0.93	0.94	0.72	0.72	0.06	0.78
	One-Shot	0.93	0.94	0.72	0.70	0.08	0.79
	Two-Shot	**0.97**	**0.98**	0.67	0.68	0.08	0.80
	Three-Shot	**0.97**	0.97	0.72	0.72	0.08	0.80
GPT-3.5-Turbo	Zero-Shot	0.75	0.78	0.43	0.42	0.05	0.76
	One-Shot	0.85	0.87	0.52	0.52	0.07	0.78
	Two-Shot	0.79	0.79	0.42	0.40	0.08	0.79
	Three-Shot	0.78	0.78	0.52	0.52	0.08	0.79
	Fine-Tuned	**0.97**	0.97	0.73	0.71	0.07	**0.85**
Mistral-7B-Instruct	Zero-Shot	0.83	0.86	0.62	0.62	0.06	0.77
	One-Shot	0.69	0.64	0.62	0.62	0.07	0.79
	Two-Shot	0.64	0.58	0.50	0.53	0.07	0.79
	Three-Shot	0.82	0.84	0.59	0.61	0.07	0.80
	Fine-Tuned	**0.97**	**0.98**	**0.75**	**0.74**	**0.10**	0.81
Llama-2-7B-Chat	Zero-Shot	0.68	0.79	0.25	0.11	0.02	0.70
	One-Shot	0.60	0.75	0.25	0.11	0.06	0.76
	Two-Shot	0.60	0.75	0.24	0.10	0.06	0.78
	Three-Shot	0.77	0.80	0.24	0.10	0.06	0.79
	Fine-Tuned	**0.97**	**0.98**	0.74	0.73	0.08	0.80
Falcon-7B-Instruct	Zero-Shot	0.59	0.75	0.46	0.46	0.07	0.78
	One-Shot	0.59	0.73	0.23	0.29	0.04	0.73
	Two-Shot	0.60	0.75	0.16	0.13	0.05	0.74
	Three-Shot	0.60	0.75	0.16	0.12	0.04	0.74
	Fine-Tuned	0.96	0.97	0.73	0.72	0.09	0.78

Table 7.2: Performance for three open and two closed models on two classification tasks and one text summarization task. GPT-4 outperforms other models in few-shot settings, but open models are competitive after fine-tuning with modest assumptions.

records annotated by three humans (Krippendorff's α of 0.88, 95% CI: 0.85, 0.90 [466]).

2. **Fact-Checking**: I use the Climate-FEVER dataset [116] to evaluate performance on a fact-checking task. Given a climate-related claim and an associated piece of evidence, the model answers whether the evidence Supports, Refutes, or provides insufficient information to support or refute the claim [116]. For predefined training, validation, and test splits, I use the version of this dataset available at `https://huggingface.co/datasets/amandakonet/climate_fever_adopted`, used in fine-tuning in-domain climate fact-checking models like a Climate-BERT [420].

3. **Clinical Dialogue Summarization**: I use the MTS-Dialog dataset [27] to evaluate models on clinical dialogue summarization, following prior work [412, 163]. Given a dialogue between doctor and patient, plus the topic (*e.g.,* medication history, chief complaint), the model must summarize the dialogue, capturing information relevant to the topic.

A simple postprocessing script removed extra words so model output could be measured against labels for tasks 1-2. Given "The answer is Supports" for fact-checking, the script removes "The answer is."

7.4.2 Performance — Fine-tuned Open Models Can Outperform Closed Models

As shown in Table 7.2, GPT-4-Turbo outperforms open models in the few-shot setting, and by substantial margins for the entity resolution and fact-checking tasks. Of the open models, only Mistral-7B-Instruct is competitive with GPT-3.5-Turbo in the few-shot setting. Fine-tuning for a single dataset epoch, however, yields open models that are competitive and in some cases even outperform GPT-4-Turbo and fine-tuned GPT-3.5-Turbo. LLaMA-2-7B-Chat achieves no more than 25% accuracy on the fact-checking task in any few-shot setting, yet outperforms GPT-4-Turbo after fine-tuning. GPT-4-Turbo also achieves the best few-shot performance on medical summarization. With fine-tuning, though, Mistral-7B-Instruct

outperforms GPT-4-Turbo few shot, achieving higher BLEU score (but not higher BERT score) than GPT-3.5-Turbo, while fine-tuned LLaMA-2-7B-Chat and Falcon-7B-Instruct achieve results competitive with few-shot GPT-4-Turbo.

7.4.3 Cost Analysis — Open Models Are More Affordable

To better understand the financial cost of customizing and using open models versus using closed models out of the box, I compute the approximate cost of inference and of fine-tuning for the climate fact-checking task. For closed models, I obtain the number of input tokens in the test dataset using the tiktoken tokenizer for OpenAI models. I multiply this total by the per-token costs published by OpenAI. I omit the cost of output tokens in this computation, which I estimate to be less than 1% of the total cost of inference for the tasks. I compute cost for open models by taking the per-hour price of the cloud instance times the runtime logged to my Weights and Biases [40] account. Costs reported are consistent with billing by OpenAI and GCP. I also report runtime for open and closed models.

If laboratory-scale AI is feasible, I expect open models to be cost-competitive with closed models, and ideally more affordable. Table 7.3 shows that the few-shot cost of GPT-4-Turbo is approximately ten times that of a few-shot open model or GPT-3.5.-Turbo. The cost of fine-tuning any open model for one dataset epoch and evaluating it once ("Fine-Tuning" in Table 7.3), a process which Performance results indicate produces a superior fact-checking model to GPT-4-Turbo, is lower than the cost of running inference once using GPT-4-Turbo in the one-shot setting. The most significant savings come when using the model after fine-tuning ("Fine-Tuned" in Table 7.3). Fine-tuned open models are much less expensive than GPT-4-Turbo, and more performant than few-shot closed models.

Closed models excel on runtime. Fine-tuned GPT-3.5-Turbo is the fastest option, and ten times faster than open models. Few-shot GPT-4-Turbo requires 1.5 times as long as few-shot GPT-3.5-Turbo, but is three times as fast as open models. Measurements do not include all costs, such as purchasing persistent disk storage, static IPs, and more reliable cloud instances, but provides an empirically grounded analysis of the cost of entry to locally train and deploy

Model	Scenario	Input Tokens	1k Token Cost	Runtime Hours	Cloud Cost	Total Cost
GPT4-Turbo	Zero-Shot	260,056	$0.010	0.32	N/A	$2.60
	One-Shot	385,926	$0.010	0.34	N/A	$3.86
	Two-Shot	484,166	$0.010	0.31	N/A	$4.84
	Three-Shot	550,171	$0.010	0.32	N/A	$5.50
GPT3.5-Turbo	Zero-Shot	260,056	$0.001	0.20	N/A	$0.26
	One-Shot	385,926	$0.001	0.23	N/A	$0.39
	Two-Shot	484,166	$0.001	0.20	N/A	$0.48
	Three-Shot	550,171	$0.001	0.20	N/A	$0.55
	Fine-Tuning	260,056	$0.003	1.54	N/A	$6.60
	Fine-Tuned	260,056	$0.003	0.11	N/A	$0.78
Falcon-7B-Instruct	Zero-Shot	N/A	N/A	0.84	$0.32	$0.27
	One-Shot	N/A	N/A	1.24	$0.32	$0.40
	Two-Shot	N/A	N/A	1.37	$0.32	$0.44
	Three-Shot	N/A	N/A	1.58	$0.32	$0.50
	Fine-Tuning	N/A	N/A	9.95	$0.32	$3.18
	Fine-Tuned	N/A	N/A	0.96	$0.32	$0.31
LLaMA-2-7B-Chat	Zero-Shot	N/A	N/A	0.91	$0.32	$0.29
	One-Shot	N/A	N/A	1.27	$0.32	$0.41
	Two-Shot	N/A	N/A	1.48	$0.32	$0.47
	Three-Shot	N/A	N/A	1.64	$0.32	$0.53
	Fine-Tuning	N/A	N/A	10.92	$0.32	$3.49
	Fine-Tuned	N/A	N/A	1.08	$0.32	$0.34
Mistral-7B-Instruct	Zero-Shot	N/A	N/A	0.97	$0.32	$0.31
	One-Shot	N/A	N/A	1.10	$0.32	$0.35
	Two-Shot	N/A	N/A	1.25	$0.32	$0.40
	Three-Shot	N/A	N/A	1.37	$0.32	$0.44
	Fine-Tuning	N/A	N/A	10.90	$0.32	$3.49
	Fine-Tuned	N/A	N/A	0.92	$0.32	$0.29

Table 7.3: Open models are less costly than GPT-4-Turbo, based on costs computed using fact-checking data. The cost of fine-tuning GPT-3.5-Turbo includes 727,845 Training Tokens, billed at $0.008 per 1,000.

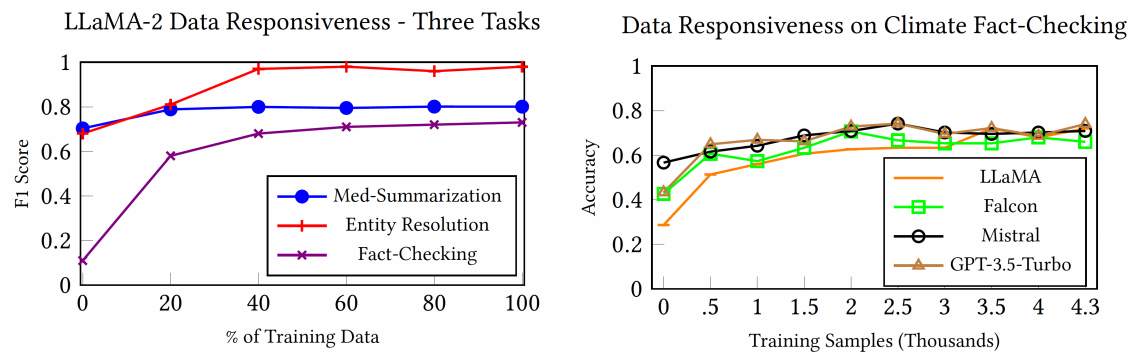


Figure 7.1: Left: Fine-tuning improvements emerge during the first 50% of the training data, only a few hundred training samples in the case of Medical Summarization and Entity Resolution. Right: Finetuned open models are competitive with finetuned GPT-3.5-Turbo with little data (1,000 fact-checking samples).

a model.

7.4.4 Data Responsiveness — Modest Fine-tuning Can Make Open Models Competitive

To understand the amount of data needed to produce a domain-specific open model, I study the performance of LLaMA-2-7B-Chat checkpoints for clinical dialogue summarization, entity resolution, and climate fact-checking tasks. I save intermediate model weights at 20%, 40%, 60%, 80%, and 100% of each task-specific training dataset, and assess the intermediate model on the full test dataset. Moreover, for the climate fact-checking task, which has a larger training set of 4,298 samples, I save checkpoints every 500 samples, and assess accuracy using these checkpoints on 150 test samples (approximately 10% of the test dataset). I save these 500-step fact-checking checkpoints for LLaMA-2-7B-Chat; Mistral-7B-Instruct; Falcon-7B-Instruct; and GPT-3.5-Turbo. Because OpenAI does not allow saving model checkpoints during fine-tuning, I submit separate fine-tuning jobs for GPT-3.5-Turbo using subsets of the training dataset.

If laboratory-scale AI is feasible, one would expect that massive data-gathering projects

would not be needed to produce a competitive in-domain model. As shown in Figure 7.1 (left plot), LLaMA-2-Chat-7B achieves BERTScore-F1 of .79 on clinical dialogue summarization after only 20% of training samples (240 samples), and .97 F1 on entity resolution after only 40% of training samples. Similarly (right plot), Mistral-7B-Instruct trained on climate fact-checking achieves accuracy of .71 after 2,000 samples, while LLaMA-2-Chat-7B achieves accuracy comparable to fine-tuned GPT-3.5-Turbo after about 3,500 samples. Fine-tuned laboratory-scale models capable of results comparable to GPT-4-Turbo can be trained using quantities of data feasible for researchers to gather. Variance among open models reflects base model benchmark performance, with Mistral generally outperforming LLaMA-2, and LLaMA-2 outperforming Falcon [223, 169], suggesting pretraining disparities (*e.g.*, LLaMA-2 pretraining on a larger dataset than Falcon) carry over during domain adaptation.

7.4.5 Model Generality — Fine-tuning Does Not Inhibit the Generality of Open Models

While fine-tuning may improve the performance of a chat-based model for a specific task, it is not clear whether this would compromise the model's general-purpose utility when a user interacts with it via natural language. To study whether the model maintains this utility, I evaluate each of the domain-specific (entity resolution, fact-checking, and clinical dialogue summarization) LLaMA-2-Chat-7B models on the other tasks for which the model was not fine-tuned. I then compare the domain-specific model's performance on each task against the general-purpose base LLaMA-2.

If the model maintains a general purpose utility, one would expect to see at worst insignificant decreases in the performance of a fine-tuned model when compared with the base model. As illustrated in Figure 7.2, performance actually increases marginally in most cases when using fine-tuned models on tasks for which they were not fine-tuned. For example, the fine-tuned fact-checking model exceeds base model performance in the one, two, and three shot settings for the entity resolution task. This may not mean that low-rank fine-tuning will always improve performance on related tasks, but the findings suggest that fine-tuning for a specific domain does not degrade the general-purpose utility of an open model.

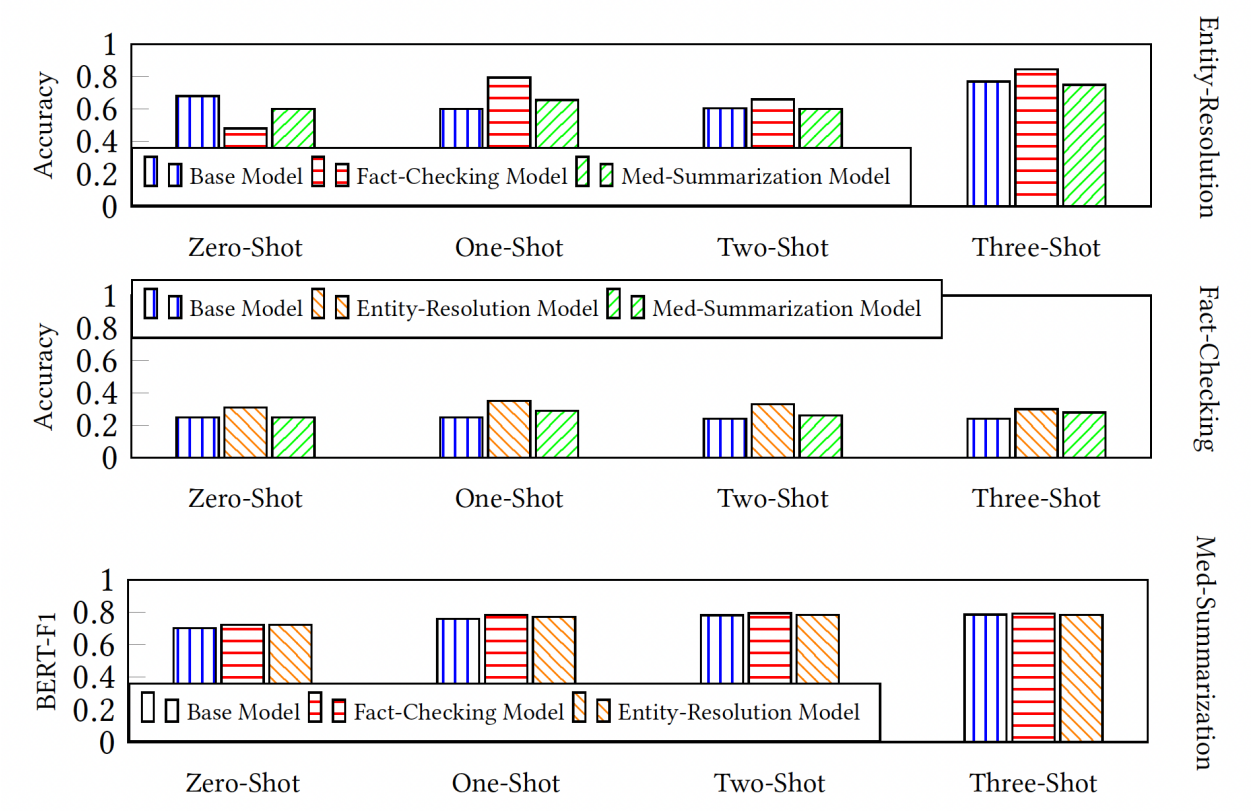


Figure 7.2: Models fine-tuned on a task using qLoRA offer strong zero-shot performance on other tasks, often stronger than the base model.

7.5 Responsible Use of Open Models

One of the presumed advantages offered by closed models is the process used to mitigate bias and prevent the closed model from generating harmful or inaccurate output. I thus evaluate three scenarios related to responsible and transparent model use: question answering under differential privacy (privacy), toxicity classification (bias), and abstention, referring to a model refusing to confidently answer questions for which it does not have the answer (transparency).

7.5.1 Differential Privacy — Privately Fine-tuned Open Models Approach Non-Private Performance

Differentially private (DP) deep learning (using a privatized gradient descent optimizer [1]) has been adopted to protect users and avoid legal risks of sensitive data use [149, 146]. While challenging in the context of language models [303], recent work [457, 448] demonstrates the potential to train general purpose models using differentially private fine-tuning on sensitive data [456]. I adopt the perspective of a small medical lab with sensitive data, seeking to privately fine-tune an open-source, general purpose medical model. I use the MedQA [193] task as a proxy for this scenario (included in the MultiMedBench [402] benchmark), simplifying it to a binary classification task. I employ private fine-tuning with qLoRA, and report results at five levels of privacy (ϵ = 0.5, 1, 5, 20, ∞, where lower ϵ denotes greater privacy, and $\epsilon = \infty$ is non-private).

Table 7.4 illustrates how challenging MedQA-TF proved for open models, which performed much lower than the state-of-the-art [402]. However, the results show that private fine-tuning allowed a model like Mistral-7B-Instruct to approach its non-privately fine-tuned performance at ϵ=20. Figure 7.3 demonstrates how different privacy settings used in Mistral-7B-Instruct impact evaluation loss curves, showing that at lower ϵ, models take longer to converge. A challenge of noisy, privatized updates is that batch size needs to be large, posing issues for lab-scale approaches that use smaller batches.

Scenario	Model (finetuned)	Acc. at Privacy Level					F1 at Privacy Level				
		$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 20$	$\epsilon = \infty$	$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 20$	$\epsilon = \infty$
MedQA-TF	Falcon-7B-Instruct	0.47	0.51	0.52	0.53	0.52	0.35	0.36	0.36	0.36	0.51
	Llama-2-7B-Chat	0.19	0.41	0.52	0.52	0.56	0.26	0.46	0.53	0.54	0.55
	Mistral-7B-Instruct	0.57	0.58	0.59	0.59	0.65	0.53	0.55	0.56	0.59	0.65

Table 7.4: Privately tuned models can approach non-private performance at lower levels of privacy.

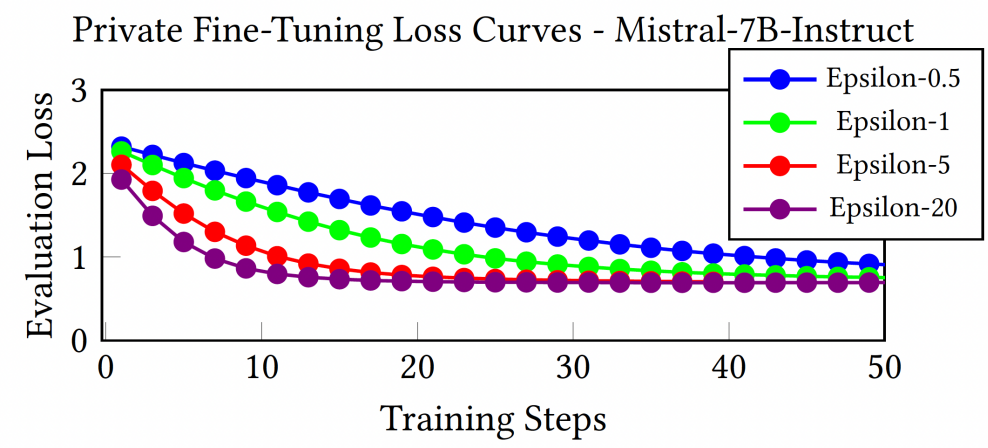


Figure 7.3: Increasing privacy (by decreasing ϵ) leads to noisier gradients, delaying convergence; but privately trained open models do learn.

7.5.2 Toxicity Bias — Open Models Improve with Fine-Tuning, But Lag Behind Closed Models

I evaluate open and closed models on a subset of CivilComments-WILDS [207], a dataset of real online comments curated from the Civil Comments platform. Dataset labels describe the toxicity of the comment and whether a demographic membership is mentioned in the comment. Models must classify whether a comment is toxic, and their classifications are analyzed through the lens of performance and fairness (whether classifications are incorrect more often for certain demographic groups). I report 1) accuracy on all comments assessed and 2) worst-group accuracy, which represents the lowest accuracy after segmenting the model's output by demographic membership and toxicity label (*e.g.,* worst-group accuracy might refer to accuracy for non-toxic comments and male demographic membership). To ensure a controlled and interpretable experiment, I limited the demographic groups to Male and Female, such that the measurements correspond to gender bias. I used 800 training, 100 validation, and 200 test samples from the dataset. Training, validation, and test data were balanced across the four groups (Male Toxic, Male Non-Toxic, Female Toxic, Female Non-Toxic).

As shown in Table 7.5, closed models outperform open models on this assessment. Fine-tuning improves overall (Mean) accuracy for Mistral and Falcon, but had no discernable effect for LLaMA-2. Fine-tuning did not increase Worst-Group accuracy over performance in the few-shot setting for any of the open models. The strongest performing model of the group was three-shot GPT-4-Turbo, which exceeded other models in both Mean and Worst-Group accuracy. Fine-tuned GPT-3.5-Turbo matches three-shot GPT-4-Turbo on overall accuracy, but not Worst-Group accuracy. However, the task is difficult, and three-shot Mistral-7B-Instruct surprisingly outperforms zero-shot GPT-4-Turbo on Worst-Group accuracy.

Scenario	GPT-4-Turbo		GPT-3.5-Turbo		Mistral-7B-Instruct		Falcon-7B-Instruct		Llama-2-7B-Chat	
	Worst	Mean	Worst	Mean	Worst	Mean	Worst	Mean	Worst	Mean
Zero-Shot	0.37	0.63	0.61	0.66	0.1	0.53	0	0.51	0.14	0.52
One-Shot	0.37	0.64	0.59	0.63	0.3	0.49	0.4	0.5	0.1	0.5
Two-Shot	0.63	0.68	0.62	0.64	0.41	0.51	0.49	0.53	0.14	0.52
Three-Shot	0.69	0.71	0.54	0.61	0.5	0.56	0.17	0.5	0.09	0.52
Fine-Tuned	-	-	0.66	0.71	0.49	0.57	0.15	0.55	0.11	0.48

Table 7.5: Fine-tuning marginally improves toxicity classification accuracy in open models, but closed models still consistently outperform them.

7.5.3 Abstention — Fine-tuned Open Models Largely Abstain from Emitting Misinformation

Instruction-tuned language models answer questions based on their parametric knowledge [290] or based on context provided as part of the prompt by the user [352]. If the model has the necessary information in neither its parametric knowledge nor the user-provided context, the model should *abstain* from answering to avoid misinforming a user [426].

Scenario	GPT-4-Turbo	GPT-3.5-Turbo	Mistral-7B-Instruct	Falcon-7B-Instruct	Llama-2-7B-Chat
Zero-Shot	0.59	0.54	0.36	0.26	0.11
One-Shot	0.56	0.46	0.33	0.31	0.11
Two-Shot	0.60	0.34	0.37	0.21	0.12
Three-Shot	0.67	0.47	0.41	0.16	0.13
Fine-Tuned	-	0.74	0.52	0.45	0.47

Table 7.6: Fine-tuning significantly improves the performance of open models on the QASPER science question answering dataset, though open models still lag behind few-shot GPT-4-Turbo and finetuned GPT-3.5-Turbo.

I evaluate the ability of open models to abstain by adapting questions from context-dependent scientific knowledge benchmarks, where some questions are designed to be unanswerable if annotators cannot find the answers based on the context. I use the *full* training set from QASPER [101] science question answering dataset to finetune and use the *answerable*

questions from the test set to assess abstention in the following way: I remove the context completely, such that the correct answer is to abstain ("Without Context" in Table 7.7). I use abstention rate to evaluate models' abstention performance following previous work [426]. Ideally, the abstention rate should be 1 if I remove the context completely. In addition to abstention, I evaluate model performance on the full QASPER test set (via F1 score) to assess tradeoffs between overall performance and abstention ability (Table 7.6). I follow the original split of train, validation, and test sets, resulting in 2,593, 1,005, and 1,451 questions respectively. Table 7.6 describes task performance on QASPER test set. GPT-4-Turbo excels in few-shot settings. Fine-tuning significantly improves task performance across models, and fine-tuned GPT-3.5-Turbo achieves the highest F1 of 0.74, 0.07 higher than GPT-4-Turbo. Fine-tuning improves Mistral-7B-Instruct, Falcon-7B-Instruct and LLaMA-2-7B-Chat, but performance does not approach GPT-4-Turbo on this challenging task.

Scenario	Model	Without Context
Zero-Shot	GPT3.5-Turbo	0.93
	Falcon-7B-Instruct	0.02
	Llama-2-7B-Chat	0.00
	Mistral-7B-Instruct	0.70
Fine-Tuned	GPT3.5-Turbo	0.53
	Falcon-7B-Instruct	0.65
	Llama-2-7B-Chat	0.99
	Mistral-7B-Instruct	0.38

Table 7.7: Without context, models that abstain well in the zero-shot setting (GPT3.5 and Mistral) do not abstain well after finetuning. Models that abstain poorly in the zero-shot setting (Falcon and Llama) *improve* after finetuning.

Table 7.7 describes results for the abstention task ("Without Context" means the model is not provided enough information to answer the question and should always abstain) using answerable questions from QASPER test set. Surprisingly, with fine-tuning, abstention

performance is reduced for the best question-answering models, suggesting an "overconfidence" effect: Models that are capable of abstaining in the zero-shot setting (GPT3.5Turbo at 0.93 and Mistral-7B-Instruct 0.70) are less likely to abstain in the fine-tuned setting (GPT3.5Turbo at 0.53 and Mistral-7B-Instruct 0.38). However, for models that are *unable* to abstain in the zero-shot setting (Falcon-7B-Instruct at 0.02 and Llama-2-7B-Chat at 0.00), fine-tuning significantly improves this capability (Falcon-7B-Instruct at 0.65 and Llama-2-7B-Chat at 0.99). Results suggest a sweet spot in balancing overall performance with the ability to abstain using ordinary training regimes.

7.6 Discussion

7.6.1 The Viability and Implications of Laboratory-Scale AI.

This research provides empirical support for the viability of adopting a "laboratory-scale" approach to AI that prioritizes user autonomy, privacy, fairness, and transparency while maintaining much of the performance and usability offered by industry-dominant corporate models. With a small GPU card, users can create domain-specific, chat-based language models and deploy them without losing the general-purpose utility and interface that makes such technologies appealing. The laboratory-scale approach intends to address, in a limited capacity, the challenges posed by scholars such as Bender et al. (2021) [28], who highlight the dangers of training language models on poorly specified web scraped data generally unrelated to the tasks for which the model will be used; Birhane et al. (2022) [44], who describe the performance-centric "values encoded in machine learning research," and highlight the field's capture by big tech companies; and Palmer et al. (2024) [293], who contend that scientists and academic researchers must justify the use of proprietary, closed models over open models. Laboratory-scale AI centers the domain-specific, responsible application of small, open models, presenting an option for scientists and public interest technologists who have good reason to avoid closed models that cannot be accessed except via a call to an API.

7.6.2 Affordances and Challenges of Open Models.

I used the libraries and model ecosystem provided by HuggingFace [432]. The Supervised Fine-Tuning trainer class provided by the TRL library made adapting open language models relatively simple, and primarily dependent on the organization of the data. The Huggingface ecosystem also supports qLoRA [107], which made customizing quantized models relatively straightforward. However, I nonetheless encountered difficulties with using open models that bear discussion. The most intractable problem I encountered in fine-tuning my own models lay in the difficulty of obtaining cloud instances equipped with even low-cost GPU hardware. I experienced consistent difficulties obtaining results due to lack of available cloud resources. Moreover, I did not expect the quantized open models I tested to run so much more slowly than the closed models I tested. This is related in part to the choice of a low-end GPU, but where inference speed makes a difference, the evidence suggests that cost-efficient, laboratory-scale models still trail closed models.

Open models showed room for improvement on tasks related to responsible use and deployment. While results on differentially private question answering show the potential for privacy-centering open models, they are impeded by small batch sizes required to use low-cost hardware. Fine-tuning has mixed effects for abstention: where a model exhibits strong question answering performance, it is less likely to abstain when it should; but when it exhibits weak question answering performance, it more reliably abstains from answering a question when it should not. Results on the toxicity bias task suggest that open models lag behind closed models on bias mitigation. Though tempting to conclude that the RLHF process used by OpenAI is the right way to address this problem, I note that LLaMA-2-Chat-7B also undergoes RLHF [399], and performs most poorly of any of the models assessed. Future research can contribute by centering these issues.

7.6.3 Limitations and Future Work.

While the work attempts to provide an open, low-cost approach, I acknowledge that open models have undergone expensive, resource intensive pretraining on large-scale, sometimes opaque datasets. While libraries like qLoRA help to enable adaptations of pretrained models, they cannot equip one with a means of circumventing pretraining, which at this time remains the only reliable means of producing a fluent, general-purpose base model. Future work might explore alternatives that change the pretraining paradigm. I also acknowledge that results from closed models may not be reproducible, should OpenAI change or remove models from its API, potentially without notifying the end user. This is a limitation of closed models that motivates the study, but also necessarily a limitation of the work. Finally, I could not reliably model the carbon cost of closed models due to uncertainties about the exact hardware used to run these models, the location of the data centers on which they run, and practices such as batching user inputs, which may allow for economies of scale.

7.7 Conclusion

The findings of this study demonstrate that choosing open models to reduce epistemic risk in consequential settings (such as those that require reproducibility and data privacy) is both possible and increasingly practical. The study showed that small, open, models are competitive with closed models, in that they are cost-efficient, responsive to user data, and robust to fine-tuning. As a methodological approach, laboratory-scale AI can serve as a basis for future scientific and public interest work, enabling practitioners to customize models without needing to rely on closed, API-based AI.

Chapter 8

TRADEOFFS OF TRANSPARENCY: DEVELOPING A HUMAN-CENTERED APPROACH TO EPISTEMIC RISK IN OPEN AND PROPRIETARY MODELS

8.1 Preface

Though my study of Laboratory-Scale AI establishes that fine-tuning small, open models can render them competitive with closed, proprietary models, it does not take into account how humans reason about the use of open models in real-world data science work, and how their interactions with open vs. proprietary models might impact epistemic risk. In this study, I sought to better understand the societal implications and tradeoffs of open models by conducting a follow-up study of their use at fact-checking organizations. The findings of this study ultimately suggest that the choice of open or proprietary is highly contextual: most fact-checking organizations use a mix of both, considering factors such as the quality of information received from the model, whether using a model would impact end user trust, and the need to keep confidential information gathered during investigations. This study continues the inquiry opened in chapter 5 into how epistemic risk is distributed across organizational infrastructures, probing into the specific data science pipelines at fact-checking organizations that incorporate open and proprietary generative models. The findings of this chapter support **Thesis Statement 6**: *In practice, human-centered data science work benefits from a blend of open and proprietary models that leverages the strengths of each paradigm to reduce epistemic risk.*

8.2 Introduction

Generative AI models have rapidly become a component of organizational infrastructure, with more than 90% of Fortune 500 companies now using ChatGPT [309]. Such models promise to transform information work by providing approachable conversational interfaces for performing complex tasks involving large quantities of text and data [166, 127]. Recent research indicates that organizational integration of generative AI can complement the skills of educated professionals, especially early in their careers, increasing productivity and job satisfaction by automating repetitive tasks and making know-how of experienced workers more available to entry-level staff [267, 59].

Despite this potential for positive impact, however, many scholars have voiced concerns over the growing reliance on closed, proprietary models [44]. Concerns have arisen from scholars in both Natural Language Processing (NLP) and the social sciences [342, 274], responding to a growing body of research contending that ChatGPT and similar proprietary models can be used as a substitute for human subjects, both for labeling data in scientific studies [147, 8], and simulating the behavior of human subjects [298, 362], ignoring the paucity of technical information available about proprietary models and the uncertain reproducibility of results obtained. Palmer et al. (2024) [293] contend that academic researchers should prioritize the use of *open models* for which the weights are available for download, and training data is specified to the end user, unless they can provide an explicit, study-specific justification for choosing a proprietary model (*e.g.*, studying the impact of OpenAI's DALL-E models on artists due to their widespread adoption [192]).

While these studies address the importance of open models for scientific integrity, they do not consider the impact that choosing open models can have on *organizations* that adopt generative language models as technological infrastructure. In the present work, I seek to better understand the societal implications of open models by studying their use at fact-checking organizations, a group that shares several characteristics that render them worthy of consideration in this context. First, fact-checking organizations routinely employ

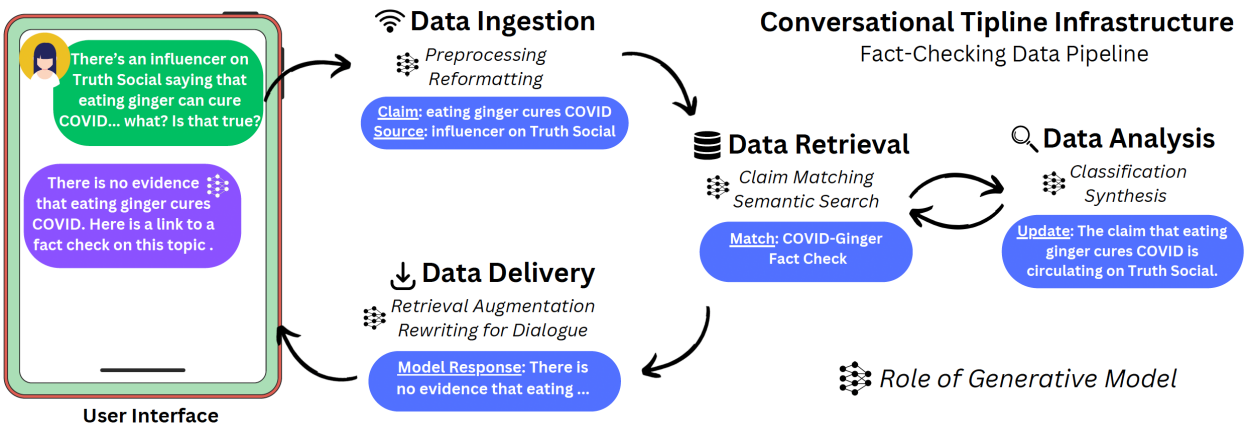


Figure 8.1: Conversational tiplines are novel data science pipelines for fact-checking accelerated by the advent of chat-based language models. Four components leverage generative AI: Data Ingestion, Data Retrieval, Data Analysis, and Data Delivery.

state-of-the-art language models in their work, lest they find themselves overwhelmed by large volumes of misinformation [100, 160]. Second, they must ensure the reproducibility, reliability, and impartiality of their work, or they will compromise both trust with their audiences and their membership in organizations such as the International Fact-Checking Network (IFCN) [310, 414]. And third, they play a vital role in maintaining the health of information ecosystems around the world [219]. Understanding use of open models at fact-checking organizations can thus provide insight into the experiences of public interest organizations leveraging generative AI in an impactful sociotechnical context. In this work, I address three research questions:

- **RQ1**: Where do fact-checking organizations employ generative AI models in their data science pipelines?
- **RQ2**: What motivates the adoption of open models by fact-checking organizations?
- **RQ3**: What prevents fact-checking organizations from further employing open models in their work?

To answer these questions, I conducted an interview study with N=24 professionals working at 20 fact-checking organizations across six continents. Adopting a human-centered approach to contextualize fact-checker perspectives on generative AI within the context of its use by practitioners, I found that fact-checking organizations reported employing generative models for Data Ingestion, Data Analysis, Data Retrieval, Data Delivery, and Data Sharing. Most participants preferred open models over proprietary models due to concerns related to organizational autonomy, data privacy and ownership, application and domain specificity, and model capability transparency. However, with a few exceptions, their use of open generative models was largely aspirational, as participants cited significant perceived shortcomings in the performance, usability, and safety of open models, as well as opportunity costs associated with not participating in emerging generative AI ecosystems offered by companies like OpenAI and Google. I make three contributions:

- **I offer a five-component conceptual model to describe where fact-checker organizations employ generative models in sociotechnical fact-checking data science pipelines.** I offer two concrete examples of in-use pipelines that employ generative models: media monitoring pipelines, and retrieval-augmented conversational tiplines (illustrated in Figure 8.1, the latter of which has seen significant improvements since the advent of general-purpose conversational models such as ChatGPT.

- **I offer taxonomies of 1) motivations of fact-checking organizations for preferring open models, and 2) limitations that prevent further adoption of open models.** I contextualize motivations and limitations by identifying the components of the data science pipeline wherein participants most located their impact, providing a grounded view of the relationship between the model itself and its organizational and societal impacts.

- **I propose a research agenda for addressing the concerns of fact-checking organizations with both open and proprietary generative models.** I offer

concrete suggestions for research addressing the performance, usability, and safety of open models, which I suggest can help further their adoption. Given that general-purpose performance of proprietary models will mostly exceed open models, and the revenue of fact-checking organizations may be dependent on producing custom models that integrate with proprietary ecosystems, I also offer directions for research addressing transparency, agency, privacy, and specificity in proprietary models.

Rather than offering a prescriptive approach to open models or an empirical study of their effectiveness, I contribute an understanding of open models in a consequential setting, including how open and proprietary models are valued in practice. I believe these insights can inform the perspectives and research agenda of the AI ethics community.

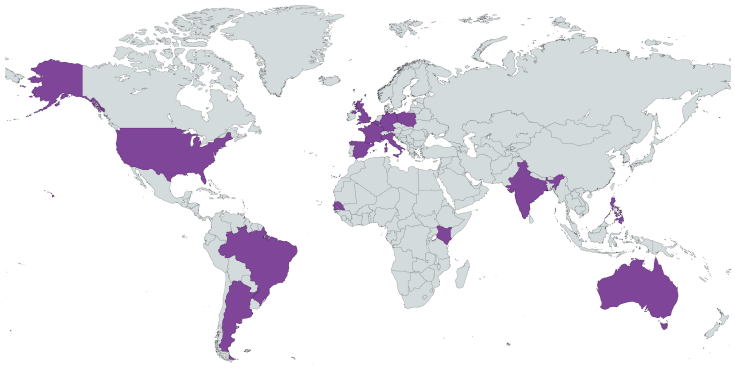
20 Organizations in 15 Countries on 6 Continents

Australian Associated Press [315], Africa Check [77], Aos Fatos [131], Chequeado [83], Code for Africa [135], Der Spiegel [369], Factly [130], India Today [395], Lead Stories [375], logically.ai [229], Maldita.es [236], Meedan [241], MindaNews [249], Newtral [264], Pagella Politica [307], Pravda [314], Rappler [333], Science Feedback [132], Snopes [367]

Table 8.1: Participants in 15 countries, 20 organizations.

8.3 Approach

I conducted an interview study with N=24 professionals at the 20 organizations shown in Table 8.1 to better understand the use of open models at fact-checking organizations. The study was approved by my university's IRB.

8.3.1 Participants

I reached out to 92 organizations via cold email, explaining the research and asking for an interview. I employed primarily purposive sampling [129] in emailing member organizations of the International Fact Checking Network (IFCN) [312] and their partner organizations, and snowball sampling [261] when individuals at these organizations offered to connect me with another organization well-positioned for participation. Five participants at five organizations enrolled in the study as a result of snowball sampling; the rest enrolled as a result of purposive sampling. Individuals at ten additional fact-checking organizations responded to my emails but lacked technical knowledge needed to respond to the questions about open and proprietary models, as their roles were related to editing or upper management. I thus excluded them from this study. Most participants were engineers, research scientists, or department managers, with experience ranging from two years to 18 years in their current role. I refer to participants using a randomly assigned number between 1 and N (*e.g.,* P24 said "...").

8.3.2 Interview Process

I developed a semi-structured interview protocol that asked participants about their organization's use of generative language models; the opportunities and challenges of generative AI in fact-checking; their organization's use of open models, and their motivations for adopting them; their reasons for using proprietary models; and what research could support the use of language models in fact-checking work. Where participants raised topics germane to the research but not covered in the interview protocol, I asked follow-up questions; for example, I asked P4 clarifying questions about their organization's beta release of a generative chatbot

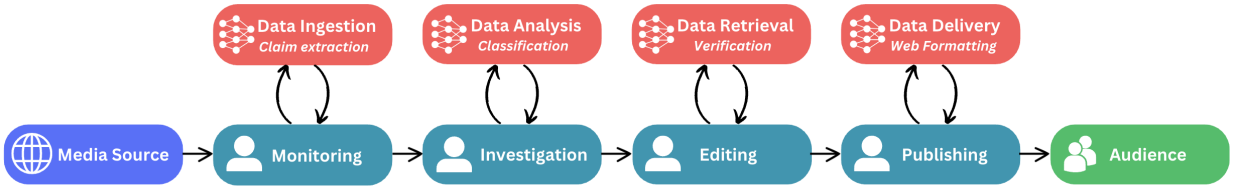


Figure 8.2: A sociotechnical media monitoring pipeline, with generative AI in red, and human processes in teal.

to collect misinformation circulating on platforms such as WhatsApp [429]. Interviews lasted between twenty-five and ninety minutes and averaged approximately forty-five minutes. All interviews were conducted in English.

8.3.3 Data Analysis

Interviews were recorded over Zoom [484], transcribed using Rev Max AI [338], and manually corrected as necessary. To answer the study's research questions, my co-author and I adopted a deductive-inductive approach to coding the interview transcripts. We employed the following deductive codes: Uses of Open Models, Motivations for Using Open Models, Limitations of Open Models, Motivations for Using Proprietary Models, and Implications for Research. My co-author and I first coded four transcripts, and I created an initial codebook that included inductively generated themes. My co-author reviewed the codebook, and we jointly revised the codes. We then coded four additional transcripts at a time until all transcripts had been coded, reviewing and revising the codebook after each round of coding. Finally, we followed a thematic analysis process [55] to generate themes that answered the study's research questions, using shared memos to precisely define the themes.

8.4 Data Pipelines in Fact-Checking

Consistent with prior work in human-centered data science, I found during thematic analysis that understanding the motivations of fact-checking organizations for using open or proprietary

models requires understanding the ways in which they collect, analyze, and exchange fact-checking data—and specifically where they use generative models in these processes. To that end, I begin by providing a conceptual model of five components of fact-checking data science pipelines in which participants described using generative models. I assign each component an icon subsequently used in the Motivations and Limitations sections to associate participant perspectives with components of the pipeline.

8.4.1 Data Ingestion

Participants reported using generative AI to collect and preprocess data, whether via *media monitoring* efforts employing AI-driven tools designed by social media companies or by the organizations themselves, or via *tipline interfaces* wherein a user can submit misinformation for fact-checking. I refer to this component of the data science pipeline as **Data Ingestion**, and denote it using an RSS icon to suggest the role of monitoring novel information.

Most participants described media monitoring pipelines like that illustrated in Figure 8.2 as an essential means of observing circulating misinformation. P3 noted that "social media listening is the main point of entry" to their data pipeline, while P11 said that they focus on monitoring WhatsApp because "the coverage is so massive" in their country. P17 said that automated approaches including generative AI were necessary for media monitoring "given the volume of production and how much content we can reasonably digest." P2 noted that one of their primary uses of generative AI was automating the data ingestion stages of their data pipeline: "we're totally focused on this pipeline, and we're capable of automating the monitoring and the detection phases." Participants also used generative AI to preprocess data for other pipeline stages. P16 said that they use generative AI to "clean up data" and that "those things [language models] are a time saver" when ingesting data. P19 described using generative AI such that "content can be synthesized and reformatted" for analysis.

Participants also described gathering information using conversational tiplines (*i.e.,* chat-bot interfaces) via which audiences can submit potential misinformation. Such interfaces are novel in comparison to media monitoring: P24 described an internal conversational interface

that existed as early as 2016, but this interface was never used for data collection. P11 said they first developed a user-facing tipline in 2019, and that they had to significantly scale the tipline during the COVID-19 pandemic, as user interactions increased more than tenfold. Conversational models utilizing modern generative AI first saw release in 2023, as P4 released a beta for a tipline utilizing ChatGPT for its back end, and P11 improved their existing framework with generative models. Generative tiplines serve several purposes, including engaging audiences and bringing in data from sources not easily observed through media monitoring. P24 noted conversational tiplines help to observe "especially provincial media that we are not that aware of or that we're not looking at regularly," while P14 noted they use a tipline to bring in new claims for investigation. P20 expressed interest in adapting their existing tipline to utilize generative AI for "constraining the conversation with [the user] to elicit more data that's actionable."

8.4.2 🔍 *Data Analysis*

Participants reported using generative AI to analyze large volumes of potentially misinforming content, leveraging its capabilities to parse highly contextual information, and utilizing few-shot approaches to avoid fine-tuning additional models. I denote 🔍 **Data Analysis** with a magnifying glass 🔍 to suggest the closer study of information.

Participants reported using generative AI to support *classification* and *synthesis* of text and multimodal content. P4 used generative AI to classify text based on constructs like urgency that they previously captured using proxies like message formatting: "instead of counting how many exclamation points a post has or how many caps it uses. . . Gen AI. . . give[s] us a score from zero to a hundred of the sense of urgency." P15 noted using generative models as part of an ensemble of deep learning and rule-based NLP classifiers. P2 used generative AI to extract and classify "patterns of manipulative messages. . . like the government doesn't want you to know this or share it," noting that generative AI can be "something like an anti-spam filter" for misinformation. P3 said they use generative AI for classifying multimodal misinformation, including memes: "the text itself isn't disinformation. The image without the

text isn't disinformation. The image plus the text can feed very clearly into a disinformation narrative." P2 noted that user-friendly generative AI enables less technical fact checkers to create classifiers: "Generative AI... democratizes who can work with AI... with an API and a little magic with a prompt, you can have something really powerful."

Participants also embraced uses of generative AI for synthesizing data. P22 said that "one of the values of generative AI is really synthesis, and looking and combing through tons of material, which... [we] will not have time for." P1 noted using ChatGPT to synthesize hundreds of documents collected every day via media monitoring, and to structure the data "in a tabular format... in 90% of the cases, it gives me a nice table." P3 used generative AI to synthesize component claims into narratives, improving the scalability of fact-checking work: "large language models are super good at basically clustering claims into a narrative. Fact checking just individual claims is whack-a-mole, a losing proposition... you'll never be able to scale it." P11 demoed a GPT-4 driven narrative system for us, explaining that it synthesizes new claims into overarching narratives already associated with fact checks, pending human review: "this summary is linked to these four different contexts that we have already received... generative AI here has proposed to us something that is a bit overarching... [it] is seeing what we produce, and the evidence that accompanies [our] debunks, and is proposing to us [a fact check] that has already been verified by a human."

8.4.3 🗄 *Data Retrieval*

Participants reported using generative AI to facilitate 🗄 **Data Retrieval** from catalogues of past fact checks or other verified sources of information maintained by the organization. I denote Data Retrieval with a database 🗄 to suggest the retrieval of stored data.

Many participants described using *Retrieval-Augmented Generation* (RAG) [216] to allow GPT models to incorporate factual data. P4 described "a RAG pipeline that connects OpenAI's GPT-4 with our database of fact checks... we have all the articles and fact checks that we ever published stored as embeddings. And then... we perform semantic search using cosine similarity, and we take the most relevant results." P19 described retrieving data

from a catalogue of past fact checks to support fact-checker investigations. P13 noted using third-party reliability metrics to determine what external content can be accessed using RAG. P11 noted that RAG allowed them "to go beyond very basic keyword searches... so that the search in the database was actually fruitful and accurate." Some participants adopted more complex methods. P22 described creating an internal knowledge graph from which generative models could retrieve content: "We started building our own knowledge graph, our own ontology, and using that, the structured data from that, to generate content." P23 created custom GPTs to retrieve data: "we built it on the claim review database... only our factcheck articles... so interacting with that search persona would give results only from the database along with a source link." P23 also produced custom models to retrieve external data: "Parliament data is completely public, so we had the tech team... scrape the entire database ... and the persona only picks up responses from this dataset."

8.4.4 *Data Delivery*

Participants reported using generative AI to support **Data Delivery** to end users on websites or social media, as well as by providing automatic responses to users via conversational tiplines. I denote Data Delivery with a download icon to suggest the transfer of data to the end user.

Among the most common uses of generative AI reported was to format content or generate metadata prior to sharing it with audiences. P1 described using GPT models to "generate hashtags for... mini FactCheck videos that we publish on TikTok." P8 said that "in most daily use cases, in terms of generative AI use, I would say it's help with promotion. So all of the social media content, coming up with summaries for SEO purposes for article publishing, title generation." P5 noted that they use a generative AI-backed tool that "suggests times the best times for us to post our content on social media based on the type of demographic our subscribers are, or our audiences are, and when they're using social media."

Participants who used conversational tiplines reported using generative AI to *deliver* information to end users, in addition to *ingesting* misinforming content. P4 said that

incorporating ChatGPT into their tipline was "an obvious use case to improve a product that was already relevant for our readers," noting their conversational tipline had "over 70,000 users, which for an organization our size is quite a lot." Several participants described an evolving information ecosystem wherein users sought information from specialized conversational agents, rather than traditional search applications. P21 said that they see generative AI as "a preferred medium for somebody to get at the work that we have done... it is summarizing or reporting on work that was done by the trusted fact checkers." P24 said that they use a WhatsApp chatbot that "allows us to answer to a high volume of messages," noting that "if you could actually ask the [chatbot], can you please tell me what the inflation was in the last five years? And it could actually answer you with information that comes from a reliable source, which we know is one of the big problems of generative AI, we think that's an enormous leap forward in the way that we can actually reach people with verified information." P19 contended that conversational agents can assist users "even if we don't have a fact check... explain this persuasion technique that's being used or the trope that is being repeated." P9 envisioned reaching younger audiences with an in-progress tool allowing a user to "chat with our archive... this is a hurdle for young people to get into the discourse, that some background knowledge is missing. Maybe the AI could help."

8.4.5 *Data Sharing*

Participants noted that generative AI assists in **Data Sharing** between organizations, in that they structure data for sharing, and serve as shared computational infrastructure. I denote Data Sharing with a Users icon suggesting transfer of data between organizations.

P18 described generative AI as a tool for exchanging fact check data and collectively scaling audiences to address problematic information: "especially during elections, we try at (anonymized) to bring together other fact checkers so that people are not working in their own little silos... that's one area in which generative AI can really help. If fact-checkers are working together, whatever data they have, they can help scale the impact of their fact check to different segmented audiences that they serve." P3 described a prototype generative AI

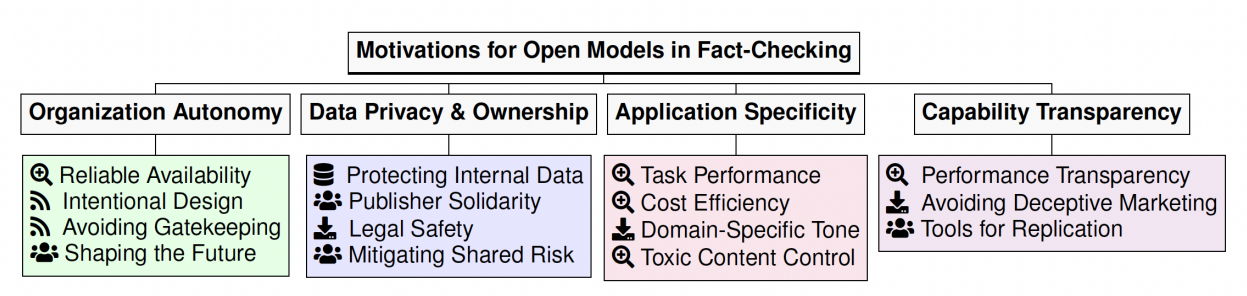


Figure 8.3: Motivations of participants for preferring open models over proprietary models in fact-checking organizations.

system operating on shared data, noting that especially in the case of elections, "fact checkers are banding together to offer a united response... European checkers will put their claims in a common database, and we will build systems in which, when we detect a new narrative... [if] it's present more than one or two countries, there will be a special task force that will be tasked with producing a debunk... we could not do it with a BERT or a Sentence BERT... early tests are promising with a [generative] large language model. This is also where the multilingual aspect comes in very handy." P11 described a similar multi-organization "project where we have installed [generative] technology... at fact-checkers in all Russia bordering countries. And we are also looking at... Spanish speaking Latin America. And we can see how common threats appear... where there are common narratives that point to particular actors."

8.5 Motivations for Open Models

I found that four primary concerns motivate the use of open models: Organizational Autonomy, Data Privacy and Ownership, Task Specificity, and Capability Transparency. I describe these concerns in turn, making reference to the components of the data pipeline with which they intersect.

8.5.1 Organizational Autonomy

Participants expressed concern that dependence on proprietary generative models could compromise the autonomy of their organizations. They noted that open models offer more **Reliable Availability** than proprietary models, which could be affected by unexpected deprecation or provider instability. Several participants described uncertainty about using OpenAI's models in the wake of its CEO's firing and reinstatement. P4 said "the whole OpenAI drama was an eye opener. If OpenAI goes bankrupt tomorrow, then it's really bad to build products that depend on their software and on their API. We'd rather just host all the models that we use." P2 said that their organization is building next-generation content moderation and misinformation detection tools to circumvent this dependence, while P7 noted that their company builds on open models, mitigating issues of deprecation: "We primarily rely on open source technologies. When Facebook releases their models and they open source it, that's what we use. We don't primarily rely on any [closed] corporate models." Participants also preferred open models because they facilitated **Intentional Design** of tools for fact-checking use cases. P8 said that proprietary tools offered by social media companies were often inadequate because they "are targeting brands. And brands like McDonald's will have certain sets of keywords that will not change over time... we want to have this additional aspect, which is discovering new keywords because we want to stay on top of narratives." P24 said "for example, social media monitoring or social listening tools... we end up developing a lot of things out of need... because we don't have the same needs as a marketing agency." Similarly, participants noted that **Avoiding Gatekeeping** by corporations motivates open models. P10 said that "we have to prove ourselves to the social media companies" to gain access to the only tools to monitor and address misinformation on their platforms, noting this process is onerous for resource-challenged local organizations. Finally, participants said open models could help fact-checking organizations in **Shaping the Future** of fact-checking. P9 argued for accelerating adoption of open generative models: "This all brings chances and holds a lot of potential for us, and we should be the one who shape this development, and not

let others be the one who dictate how we have to deal with this at some point because we waited too long." Similarly, P22 remarked that "fact-checkers need to be part of the creation of these tools... a lot of the tools that we are seeing don't really quite fit our use cases... it's about the process by which things are created."

8.5.2 Data Privacy and Ownership

Participants described concerns surrounding data privacy and ownership as a central motivation for using open models. P9 preferred using open models for **Protecting of Internal Data**, saying that "we have very sensitive material that we're working with here, investigative reporting and investigative stories, and we don't want this to be used in [corporate] models and as training material," and further noting that they use cloud instances hosted only in Europe for "a sort of legal safety" due to stricter European data protection laws. P4 said that, due to data privacy concerns, "we have as a policy to always prioritize using open source software where we can." P21 noted "I blocked [OpenAI's] crawlers from being able to train on our content until some type of commercial compensation becomes available. I think most of us are kind of waiting, holding our breaths for the New York Times case lawsuit to play out because otherwise individual orgs the size of the fact checkers, we don't really have the leverage to accomplish what that lawsuit stands to do in setting a precedent." Participants also said that **Publisher Solidarity** motivated use of open models over proprietary models that profit from journalistic organizations' data without consent. P4 noted their discomfort with "the notion that those companies are profiting using other companies' and other people's work." Participants also preferred open models that disclosed their training datasets, providing a sense of **Legal Safety**, especially for user-facing applications. P4 said that "the issue of copyright is a big one, especially for image generation... we could never use anything, any tool that generate images in our workflow, because we don't know how most models were trained." P8 noted that they delayed using generative models due to fears of copyright infringement: "With generative AI, we were... scared to use it, because of the fact that we don't want to feel like we are plagiarizing... because of the possibility of... [copying]

other articles." Participants noted that open models at least disclose their training datasets, offering some clarity concerning the risk of infringement. Finally, participants preferred open models for **Mitigating Shared Risk** when building technologies for the fact-checking community. P4 created a transcription tool for the community using OpenAI's models, but noted is not in production because "a lot of people have concerns about sending their data - interviews, important interviews - to OpenAI."

8.5.3 Application Specificity

Participants preferred open models for specific applications that demanded high performance, domain-specific tone, and control over toxic content. P14 said that their organization preferred small, fine-tuned open models for internal data analysis, noting that they exhibit stronger **Task Performance** and **Cost Efficiency** than proprietary generative models. P6 echoed this, noting that their organization still prefers thoroughly vetted, task-specific models for many tasks, despite hype about replacing these methods with proprietary models. P10 considered GPT models one of many tools, not the sole solution to any problem involving language, despite the marketing of proprietary models. Participants also noted that proprietary models intended for general-purpose use couldn't achieve **Domain-Specific Tone** for end user applications. P4 said "out of the box, GPT-4, for instance, you get very decent results, but it's also very generic." P8 noted that, as a result of RLHF, GPT-4 "sounds completely unnatural," rendering it difficult to incorporate in their user-facing applications. Participants also reported using open models to gain more granular **Toxic Content Control**. Participants including P17 took issue with corporate models' one-size-fits-all approach to alignment with human values, which they noted could hamper the model's ability to respond to toxic and hateful content that fact-checking organizations handle during their work. P5 noted that the OpenAI's RLHF process makes it difficult for models "to unlearn things that you've already taught it through human feedback," which are not advantageous for many fact-checking applications.

8.5.4 Capability Transparency

While participants agreed that GPT-4 outperformed open models, they also said that the capabilities of open models were presented with more **Performance Transparency**. P3 expressed surprise that GPT-4 performed poorly for restructuring their data, a task they thought fell within the model's capabilities, noting that while it looked reasonable "on the surface... when you actually dug into whether it was structured coherently, it wasn't as good." Participants also said that **Avoiding Deceptive Marketing** motivated use of open models in settings involving user interaction. P17 said the presentation of models like ChatGPT encouraged inappropriate trust by non-experts: "Yes, there is a popup on ChatGPT, and a line at the ending that tells you the results could be inaccurate, so beware. But how the interface is built, and how it is marketed, how it is presented, and the fact that it answers you in a confident way... there is a constant behavioral trick... that what you have in front of you, what a machine is telling you, answering to your prompt, is the truth." P1 echoed this sentiment, noting that "There's a problem with people assigning too much credibility to large language models." Finally, participants said that providing **Tools for Replication** motivated open models. P8 noted that "we want to give a reader the ability to replicate our research in total... every source, everything." P9 said that, while all models created problems of explainability, API-gated proprietary models introduced more "black box problems" than open models, rendering reproducibility uncertain for end users.

8.6 Limitations of Open Models

Despite the many motivations participants gave for preferring open models, most nonetheless used primarily proprietary models, especially OpenAI's GPT models [288], citing their Performance, Usability, Safety, and the Opportunity Costs of not participating in proprietary AI ecosystems.

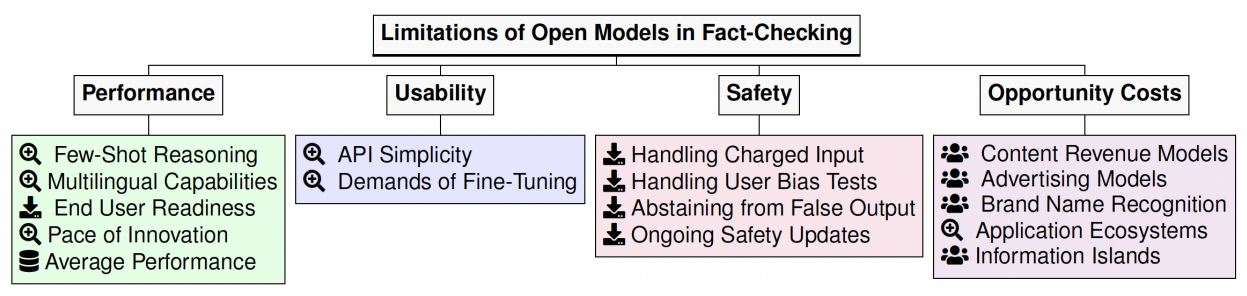


Figure 8.4: Limitations of open models described by participants as preventing their further adoption in fact-checking.

8.6.1 Performance

While fine-tuned open models may achieve the strongest performance on a given task, all participants said that GPT-4 was the best-performing *general-purpose* generative model, motivating its adoption over open models in many settings where general-purpose reasoning is preferable to task specialization. P13 noted that they use GPT-4 over LLaMA models due to **Few-Shot Reasoning** disparities, while P9 echoed this in noting that "if you want to work with a narrative generative AI model, then I think there is, at the moment, no alternative to GPT-4." Participants also noted that open models lag OpenAI significantly in **Multilingual Performance**. P4, who reported creating primarily non-English models, said that "we tried the largest [LLaMA models]... usually the models tend to perform poorly in languages that are not English. And in the case of OpenAI, it's pretty good." P4 also said that performance disparities affected the **End User Readiness** of open models: "when we were developing [our conversational tipline], we tested a bunch of models, especially LLaMA-2, but in terms of performance, it's behind OpenAI significantly... I'd love to use open source models." P2 said that the **Pace of Innovation** in generative AI made it untenable to try to build open models: "the pace of innovation nowadays is so quick that it's very difficult to keep the pace... you don't know which is the new tech that you need to use or which is the model that is going to work... even research institutions, they don't

know." In some cases, participants reported that GPT-3.5-Turbo's inexpensive **Average Performance** was good enough. P1 said that, for retrieving data internally, they needed high recall and not necessarily high accuracy, since human fact-checkers would see the data. They noted "OpenAI is... so easy and cheap. We pay a couple of dollars per month... I can use the old (anonymized) servers to run an open language model and see what it can do. But the savings would be minimal, like a couple of dollars, and the results would probably be worse."

8.6.2 Usability

Most participants said the ease of calling proprietary APIs motivated their use of proprietary generative models. P19 said that, even though they prefer to build on open source models, **API Simplicity** motivated them to OpenAI: "OpenAI models are convenient for prototyping because it's just an API call, and you don't have to worry too much about it." P2 said, "what is the good part of OpenAI? Everything with an API is much easier." However, P2 also pointed to the possibility that OpenAI's perceived edge in usability is actually due more to its market dominance, noting that "there are a lot of different frameworks now... [such as] LangChain, there are a lot of frameworks that make it easier to work with [open models], and they are the solution." Participants noted the **Demands of Fine-Tuning** open models drove them to use few-shot proprietary models. Even if task-specific performance exceeding GPT-4 is achievable with fine-tuned open models, data scientists at fact-checking organizations may not have the time to invest in fine-tuning. P2 said "the other issue for typical fact-checking organizations, is we are not Microsoft, we are not big technology companies, is the cost of the systems ... to fine tune your own language model it is too... time demanding." P4 said it was hard to find time to understand the opportunities of fine-tuning, noting they only used LLaMA "out of the box, we didn't try to fine tune it. Maybe in the future we will, but... resources are limited, so we opted not to explore more."

8.6.3 Safety

Participants said that proprietary models offer advantages for user-facing applications because of the safety features built into them by larger technology corporations. P13 noted that because user-facing fact-checking technologies typically involved **Handling Charged User Input**, reliably adhering to ethical guardrails was essential for maintaining user trust, and open models could not always accomplish this. P4 said that user-facing technologies also had to be prepared for **Handling Bias Stress Tests**, noting that users actively attempt to probe their in-beta conversational model (leveraging GPT-4) for political biases: "one of the most common types of questions that I think people were asking the bot... [was] just swapping the name of the [politician] you're asking about... and so far we haven't noticed anything that would be concerning for us... it's not symmetric, but I think it usually gives you a quite nuanced answer." P4 also noted that proprietary models tend to outperform open models in **Abstaining from False Output**, noting that users most frequently complained that their RAG-enabled GPT-4 chatbot couldn't answer a question, but that "in most cases, we didn't have the answer. So the bot behaved as expected as it should, which is to say that it doesn't know the answer to a question instead of trying to come up an invented answer." Finally, participants said that proprietary model developers were better positioned to perform **Ongoing Safety Updates**. P21 said that "fact check organizations are underequipped to really test everything and keep up with everything on an ongoing basis"; accomplishing this with open models, they said, would require "revenue streams... that's still trying to be figured out."

8.6.4 Opportunity Costs

Participants worried that using open models would entail foregone opportunities for integration into an emerging information ecosystem, potentially lessening the relevance of their content and precluding them from taking advantage of new streams of revenue. P9 said that **Content Revenue Models** would "totally define how we will work with Gen AI models,

and if we will work in close partnerships with these companies, or if we will work with open source models, or if they will be state funded projects. All the journalistic companies say we need it to save democracy, and we push for open source models; or we say, no, we're fine, we're getting millions and millions from Google, and OpenAI, and Amazon, and so on. So yeah, we're sort of at a crossroads." P21 noted that, in addition to direct compensation for content, proprietary models might be preferred due to **Advertising Models**, noting that "the value exchange that we're familiar with today with Google search is you structure your metadata the right way, and you'll show up in Search, and you'll get clicked on, and you'll generate ad revenue. And perhaps Google's even the one who buys those ads, and the revenue comes from them anyway. I think that [Google is] more ready as an organization to think about paying publishers. With ChatGPT that's not really established yet." P21 also said that they hoped to leverage the **Brand Name Recognition** of emerging AI ecosystems to increase their reach, noting "ultimately ChatGPT is the biggest brand name in chatbots, and they are also already integrated and backing Bing... to reach all of the users who are using chatbots, it's not realistic to think that we'll make the biggest splash just by having our own private code base and onsite chat experience. I do think that we have to be able to play into the bigger arena." P10 further discussed the possibility of integrating their internal models into broader **Application Ecosystems**, noting that "we use the Google workspace workflow, so it really helps to incorporate the Bard features." Finally, P21 raised broader concerns about the information ecosystem, describing the possibility of **Information Islands** if internet users privately interact with open chat models: "if people choose to search or engage with a chat bot that... isn't trained on your content, we stand to have these kind of information islands... someone who searches on Google today or Bing today, they're going to get [our content] as results that were fully enabled to be crawled by search engines. But as more and more chat bots emerge... that is kind of a threat, I think, to information integrity overall."

Research Directions for Use of **Open Models** in Fact Checking		
Concern	Research Question	Research Directions
Performance of Open Models	How can open models offer competitive performance to proprietary models while maintaining an approachable conversational interface?	Developing scale-efficient open models; Developing more suitable evaluation suites for fact-checking; Developing evaluations specifically for open models.
Usability of Open Models	What kinds of open application interfaces can help fact-checkers feel as comfortable with open models as proprietary models for inference and fine-tuning?	Creating open source or public APIs of comparable simplicity to OpenAI; Decreasing the time and expertise demands of fine-tuning.
Safety of Open Models	How can open models achieve the actual and perceived safety of proprietary models without incurring significant time and cost burdens?	Community standards for lightweight red-teaming of open models; Technologies to monitor the fairness of model responses in user-facing conversational tiplines.
Opportunity Costs of Open Models	How can open model ecosystems facilitate reliable revenue streams for fact-checking organizations similar to those in proprietary model ecosystems?	Fostering community and collaboration between fact-checking organizations and AI developers; Developing a revenue model for open model ecosystems.
Research Directions for Use of **Proprietary Models** in Fact Checking		
Lack of **Autonomy**	How can proprietary model developers assure clients of access to models integrated in data pipelines?	Developing approaches to allowing selective access to deprecated models.
Lack of **Data Privacy and Ownership**	How can proprietary model developers guarantee clients that their data won't be used inappropriately? How can they guarantee clients that using their products will not put them in legal jeopardy?	Developing models of compensation for publishers; Supporting legal standards for client data privacy; Clearly communicating when data will be retained or used outside of initial context.
Lack of **Application Specificity**	How can developers afford clients more control over tone in user-facing applications and more contextually appropriate means of processing toxic content?	Developing approaches for personalized models that meet fact-checking use cases without compromising model safety.
Lack of **Transparency**	How can developers provide transparency about model capabilities and afford fact-checking organizations the ability to explain their use to audiences?	Communicating capabilities of domain-specific GPTs through systematic quantitative & qualitative evaluation; more explainable outputs like token-level probabilities.

Table 8.2: Research directions for addressing limitations of both open and proprietary models proprietary models in fact-checking.

8.7 Discussion

These findings add new perspective to the evolving conversation about the use of open or proprietary generative AI, surfacing the tradeoffs that face *organizations* as they consider whether and where to adopt open models. Some tradeoffs would seem contradictory without being contextualized within the fact-checking data science pipeline. For example, participants chafed at limitations imposed by toxicity guardrails when using OpenAI models for internal analysis tasks, yet reported *relying* on those guardrails for user-facing applications. Moreover, while participants preferred fine-tuned open models for strong performance on internal tasks, they also accepted "good enough" performance via the cheap, easy-to-use GPT-3.5-Turbo API for high-recall data retrieval tasks, over the time cost incurred for an open model.

Table 8.2 outlines research questions and directions arising from this work. Fact-checking organizations appear at first to face diverging futures: one in which they adopt open models and control their AI infrastructure; and another operating profitably in the emerging ecosystems of chat-driven interfaces now collecting and delivering information in interactions with users. However, the findings surface that the future may resemble the present, with organizations employing specialized open models for mission-critical tasks, proprietary generalist models to handle interactions with users, and the least expensive option when performance just needs to be good enough. Even so, proprietary models face challenges around data ownership. Participants voiced unease with their work and that of their colleagues being used without consent to create a lucrative competing product. Some participants also avoided generative AI for fear of inadvertently committing plagiarism. These concerns echo those of many journalists, as well as domains like visual art [192]. Addressing data privacy, ownership, and compensation may prove paramount for proprietary model providers to establish relationships with producers of factual and creative content.

8.7.1 Limitations and Future Work

While I sought to capture perspectives of fact-checking organizations globally, the participants all spoke English, and most used models in English. Aside from noting that OpenAI's models outperform open models in multilingual settings for generalist conversational tasks, this work does not speak to the additional intricacies of multilingual NLP, including NLP in low-resource languages. Additionally, this work is concerned with open generative models, rather than with all open models. Future work might study perspectives on open models and proprietary models of any architecture.

8.8 Conclusion

Rather than offering a prescriptive approach to open models or an empirical study of their effectiveness, this study contributed an understanding of open models in a consequential setting, including how open and proprietary models are valued in practice by users who have high-stakes epistemic needs. The research shows that, while proprietary models are associated with clear and pressing sources of epistemic risk, open models introduce context-dependent risks of their own. Drawing on these findings, I offer recommendations on how the context-sensitive selection of proprietary and open models can provide organizations with an effective approach to managing epistemic risk.

Chapter 9
CONCLUSION

This dissertation has presented a series of approaches to epistemic risk in generative and general-purpose AI. It contributes approaches to measuring and understanding epistemic risk in AI; designing systems and interfaces that mitigate epistemic risk in AI; and evaluating the capabilities of open models to present a viable alternative to the heightened epistemic risk of proprietary AI. This research brings together perspectives from human-centered computing, AI fairness and ethics, and psychology and the social sciences to produce ways of contending with the risks posed by an emerging general-purpose technology, while also appreciating the opportunities offered by this technology to produce a healthier information ecosystem.

9.1 Findings and Implications

In this section, I summarize the contributions made in this dissertation, considering the potential implications of the research as generative and general-purpose AI continue to reshape our sociotechnical infrastructures.

9.1.1 Understanding Factors Contributing to Epistemic Risk

Understanding the sources of epistemic risk in modern generative and general-purpose AI systems is a complex problem due to the opacity of the enormous datasets on which they train [28], the capability of such systems to issue a response to virtually any prompt entered by a user [161], and the fundamental technical difficulty of understanding deep neural networks themselves [122], the technology underlying these systems. Though this dissertation cannot render such systems interpretable enough to fully explain epistemic risk, it does contribute approaches that increase our understanding of the sources of such risk. First, chapter 3

demonstrates that even high-quality data like that created by news organizations can produce epistemic problems when AI models are trained on this information and then used in a different context. Though the information is accurate, the qualities that make a story newsworthy may well render it inappropriate for forming an epistemically reliable representation of a social group. Second, chapter 4 provides quantitative evidence of the relationship between the scale of the data used to train a model and the unwarranted inferences made by that model. Though prior work has expressed justified concern about scale in language models and identified increases in ethical problems like hate speech as training dataset size increases [42, 28], chapter 4 presents one of the first studies to demonstrate that novel *forms* of epistemic risk (in this case, physiognomic facial impression biases) can arise as dataset size is scaled.

Though approaches that scale a model's pretraining dataset have recently run into barriers [35], such as challenges in expanding the size of existing datasets [408], or legal challenges related to the use of copyrighted data [233], recent research has found new frontiers for exploiting scale. These have included, for example, scaling inference-time compute to generate better answers to users [257, 286]. Rather than training on more information, that is, the model produces more information and then uses that information to work through its response to the user. This is one of the foundations of recent reasoning models like DeepSeek or OpenAI's o3, which produce long "thinking" chains in response to a user's input before generating a user-visible output [159, 286]. Though this technology may ultimately improve the epistemic reliability of generative models by helping them produce more reliable answers, and may also further improve the performance of small, open models in domains like mathematical reasoning [159], research has yet to probe what unintended side effects may also from these new forms of AI scaling.

9.1.2 Designing to Mitigate Epistemic Risk and Improve Information Quality

Generative and general-purpose AI present many opportunities to designers, as they can be integrated into widely differing settings and can be used to create new data or manipulate data submitted by a user [425]. This dissertation makes two primary design contributions

that intend to help mitigate epistemic risk. Chapter 6 introduces a novel design framework based on Nonviolent Communication for helping to produce technologies that help people communicate more clearly about their interpersonal needs. It also highlights the potential for AI to introduce new epistemic risks in the mediation of interpersonal communication, including empathy fog, a situation wherein an individual can no longer distinguish between true empathy communicated by a person over an online platform and AI-manufactured empathy. Chapter 5 considers how generative AI is impacting fact-checking organizations, providing a concrete array of opportunities and challenges of using generative models in a high-stakes epistemic setting. Despite the inevitable existence of epistemic risk when using a generative model, this study produced a contribution to the design space of generative models focused on the verification of information, drawing on interviews and the code of principles of the IFCN to improve the quality of information produced by fact-checkers. This suggests that design may help to produce tractable approaches for using generative models in many similar epistemically consequential settings.

The idea of using generative models to improve information quality has already become mainstream in some domains of science. Consider, for example, that the International Conference on Learning Representations (ICLR), one of the most respected venues for publishing machine learning research, has begun using generative models to provide feedback to human reviewers for the 2025 conference [410]. That is, the model assesses the quality of a human review of a conference paper and suggests ways to improve it prior to release to the paper's author. Designs that support epistemic interventions like this can help to understand whether generative models can play a role in improving the health of critical—and sometimes overwhelmed [253]—information ecosystems like scientific publishing. However, we should keep in mind that unintended consequences—like empathy fog—can occur when a technology is used to automate a task that actually requires real human attention to have meaning. Identifying where human input remains important will likely be critical for maintaining the epistemic authority of scientific institutions.

9.1.3 Evaluating the Role of Open Models in Addressing Epistemic Risk

This dissertation reckoned with the heightened epistemic risk of closed, proprietary AI models by evaluating the competitiveness of small, open models, and by probing the adoption of open vs. proprietary models by organizations with significant epistemic needs. Chapter 7 empirically evaluates small, open models, finding that in task-specific settings, they can match or even exceed the performance of proprietary models on tasks of epistemic consequence. Small, open models also present a cost-efficient alternative to larger models, though they require some technical skill to fine-tune, and their fairness and speed of response leaves much room for improvement. Chapter 8 investigates the approaches taken to the use of open vs. proprietary generative models by fact-checking organizations, finding that fact-checkers see not a binary choice but a highly contextual one, dependent on the sensitivity of the information with which they are dealing as well as their need for a reliable and reproducible output. This study also found that data privacy and ownership factor significantly into the decisions made by many fact-checking organizations with regard to their use of generative AI, suggesting the need for a better model for compensating creators of factual content.

Where open models are the most appropriate choice, one promising and now widespread technique for producing small, open models involves "distilling" the information learned by much larger models by training a smaller model on the the larger model's output [179, 174]. This technique was used to produce highly capable small reasoning models, like a 1.5-billion parameter distilled version of DeepSeek R1 [160]. Making such models open source or open weight significantly improves the opportunities available to the scientific community. Techniques for producing ever-smaller models may also enable greater personalization of AI, as it becomes more feasible to run such models locally on device [124] and potentially to update them to respond specifically to the individual using them. As discussed in greater detail below, the potential for AI-personalized information presents both epistemic opportunities and potential threats to the broader information ecosystem.

9.2 Directions for Future Work

Continuing progress in generative and general-purpose AI will demand new approaches to epistemic risk, and it may also enable solutions to some of the difficult epistemic problems facing society. In this section, I discuss potential directions for future work addressing novel risks and opportunities.

9.2.1 Epistemic Risk in AI Ecosystems

Several of the fact-checking organizations I spoke with anticipated the need to adapt their business models for the emergence of new information ecosystems like the OpenAI GPT Store [285]. Already, forward-thinking companies like Canva and Kayak have created customized GPT Store chatbots that serve as interfaces to functionality traditionally presented through a website. But how should institutions that specialize in providing information as a service configure and advertise custom conversational models? And what norms can emerging AI ecosystems establish to help organizations connect responsibly to their audiences? I scratched the surface of these questions in a recent position paper at the ACM CHI 2024 Dark Patterns Workshop [441], which contends that *misleading presentations* of model capabilities—including in such sensitive domains such as law and medicine—are common even among the highest ranked models on the GPT Store. Future work might undertake comparative studies of marketplaces for general-purpose AI (such as the GPT Store [285], the Claude Prompt Library [10], and the HuggingFace open model ecosystem [432]), or build interfaces that allow controlled studies of how to accurately communicate the capabilities of customized chatbots.

9.2.2 AI Alignment for Epistemic Reliability

Could the epistemic problems posed by AI be addressed by an alignment strategy that explicitly centers reliability? Most generative models are aligned to be "helpful" and "harmless" [20], a policy which also renders them obsequious and verbose [363]. But perhaps such models could be aligned to act in accordance not with user preferences but with scientific best

practices, and to emphasize clarity, consistency, and succinctness. Recent work introduces a promising method for this: Jang et al. (2023) [189] decompose preference alignment into multiple characteristics to which a model can be aligned, such as "expert," "unpretentious," "informative," and "concise," and allow the user to merge in preferred characteristics with an open chatbot model. This might be tested for its capability to produce model alignments that reduce formatting-based variance and inappropriate deference to users. Efficacy could be assessed in downstream tasks, measuring variance across prompts with the FormatSpread [360] and Validity Rate [6] metrics. If successful, the method could provide a more reliable option for producing user-facing chatbots in scientific research and other epistemically consequential contexts.

9.2.3 Emergent Epistemic Risk in Frontier AI Models and Systems

My studies of emergent bias in systems like CLIP have revealed that tests designed to measure bias in already existing systems like online search failed to anticipate the novel forms of epistemic risk possible in new models [437, 439]. As generative and general-purpose AI continue to improve, what novel risks will they introduce, and how can organizations develop a means of monitoring and managing these risks even as they leverage the capabilities of new models? Future work must study the challenges presented by agent-based AI systems, which utilize language models to create instructions based on a user's input and execute them using programmatic tools [298, 357]. Such systems promise new efficiencies that can save users time by performing more complex tasks—but also introduce new uncertainties about the source and validity of information.

9.2.4 AI-Personalized Information

Recent updates equip ChatGPT with what OpenAI refers to as "Memory"—in practice, a series of notes made by the model about the user, leveraged during future conversations to inform how the model communicates with the user [287]. While organizations have long used metadata to gain insight about users based on the websites they visit and the purchases they

make [339], generative AI provides a much more nuanced way to observe a user's preferences and habits. Generative models open the way for applications that help organizations deliver more targeted information; for example, a generative model might be able to rewrite a news story or a fact-check, and draw on the notes maintained in Memory to deliver a version designed to that specific user. The rewritten article might adopt an ideological stance more in line with the user's own perspective, or lead with content that speaks to the reader's interests. Future work might build interfaces to study the effects of delivering personalized information using the notes maintained by generative models. Such studies might address the question of trust in generative models as information arbiters, as well as the potentially undesirable effects of such systems on user privacy and autonomy.

9.2.5 Contextual Integrity of User Data

Generative AI possesses the capability to collect, solicit, and even make decisions about data from users, including sensitive personal data such as health information, personal financial information, and private conversations [250]. Sharing such information with a generative model might enable that model to provide a more nuanced and contextual response to the user, making this exchange seem like a reasonable proposition [475]. Yet the ultimate recipient of this data is not a chatbot but an organization serving the chatbot, and the potential cost of sharing such information could be high, were it to be used outside of its original context. Future work might consider how to equip users with a means of reasoning about the contextual integrity of their interactions with AI interfaces, and the potential costs associated with sharing sensitive personal information. It might also probe the tradeoff between privacy and user perceptions of utility or personalization, to understand how users reason about the *value* of the privacy of their data.

9.2.6 Natural Language Explanations for Detecting and Addressing Epistemic Risk

Recent research suggests that generative models can produce human-understandable explanations of other generative models. For example, Bills et al. (2023) [41] use GPT-4 to explain

why every neuron in GPT-2 activates in response to user input - and then to simulate which neurons would likely activate given a new prompt. Similarly, Ghandeharioun et al. (2024) [145] use algebraic interpretability measures [144, 277] in tandem with a 13-billion parameter Stable Vicuna chatbot [85] to explain the activation patterns of a smaller Vicuna chatbot. If a model could explain *risks* in natural language, this could inform strategies for mitigating risk in the smaller model, potentially by using the larger model to produce targeted synthetic data useful for risk mitigation. This could also allow for partial automation of searching for risks in need of mitigation.

9.3 Final Remarks

Though the problem of epistemic risk in generative and general-purpose AI remains far from solved, the contributions of this dissertation provide a means of characterizing and contending with manifestations of epistemic risk in these technologies, in some cases enabling such models to be used in the service of information integrity. These approaches and others building upon them can help to reckon with changes to the process of creating and disseminating knowledge as individuals, organizations, and even governments begin to integrate general-purpose AI more completely into the epistemic foundations of our society.

BIBLIOGRAPHY

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, , 2016. .

[2] Abubakar Abid, Maheen Farooqi, and James Y. Zou. Persistent anti-muslim bias in large language models. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021.

[3] Alberto Acerbi and Joseph M Stubbersfield. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120, 2023.

[4] Hammaad Adam, Ming Yang, Kenrick D. Cato, Ioana Baldini, Charles R. Senteio, Leo Anthony Celi, Jiaming Zeng, Moninder Singh, and Marzyeh Ghassemi. Write it like you see it: Detectable differences in clinical notes by race lead to differential model recommendations. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022.

[5] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: Towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021.

[6] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR, 2023.

[7] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, ():, 2023.

[8] Mostafa M Amin, Erik Cambria, and B Schuller. Will affective computing emerge from foundation models and general ai. *A first evaluation on ChatGPT. ArXiv, abs/2303.03186*, 2023.

[9] Lori Andrews and Hannah Bucher. Automating discrimination: Ai hiring practices and gender inequality. *Cardozo L. Rev.*, 44:145, 2022.

[10] Anthropic. Meet claude. `https://www.anthropic.com/claude`, 2024. [Accessed 17-04-2024].

[11] John Antonakis and Olaf Dalgas. Predicting elections: Child's play! *Science*, 323(5918):1183–1183, 2009.

[12] James Arnéra, Chun Hei Michael Chan, and Mauro Cherubini. Digital, analog, or hybrid: Comparing strategies to support self-reflection. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, pages 3435–3452, 2024.

[13] Jeffrey Jensen Arnett. Adolescent storm and stress, reconsidered. *American psychologist*, 54(5):317, 1999.

[14] Michael C Ashton and Kibeom Lee. Honesty-humility, the big five, and the five-factor model. *Journal of personality*, 73(5):1321–1354, 2005.

[15] Abdolghader Assarroudi, Fatemeh Heshmati Nabavi, Mohammad Reza Armat, Abbas Ebadi, and Mojtaba Vaismoradi. Directed qualitative content analysis: the description and elaboration of its underpinning methods and data analysis process. *Journal of research in nursing*, 23(1):42–55, 2018.

[16] Atlas.ti. Master your research projects with the power of ai. https://atlasti.com/, 2024. [Accessed 02-08-2024].

[17] Axel Aubrun and Joseph Grady. Aliens in the living room: how tv shapes our understanding of 'teens'. *The Frameworks Institute*, 2000.

[18] Theophilus Azungah. Qualitative research: deductive and inductive approaches to data analysis. *Qualitative research journal*, 18(4):383–400, 2018.

[19] Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105*, 2024.

[20] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, ():, 2022.

[21] Parissa J Ballard, Lindsay Till Hoyt, and Jasmine Johnson. Opportunities, challenges, and contextual supports to promote enacting maturing during adolescence. *Frontiers in Psychology*, 13:954860, 2022.

[22] Amanda Baughan, Justin Petelka, Catherine Jaekyung Yoo, Jack Lo, Shiyue Wang, Amulya Paramasivam, Ashley Zhou, and Alexis Hiniker. Someone is wrong on the internet: Having hard conversations in online spaces. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–22, 2021.

[23] Amanda Baughan, Larry Tian, Pranav Shekar, Amy Zhang, and Alexis Hiniker. Supporting hard conversations in close relationships through design. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–22, 2024.

[24] Be My Eyes. Announcing 'be my ai,' soon available for hundreds of thousands of be my eyes users. https://www.bemyeyes.com/blog/announcing-be-my-ai, 2023. [Accessed 02-29-2024].

[25] Be My Eyes. See the world together. https://www.bemyeyes.com/, 2023. [Accessed 02-29-2024].

[26] Jeff Beckman. Openai statistics 2023: Growth, users, and more. https://techreport.com/statistics/openai-statistics/, 2023. [Accessed 19-01-2024].

[27] Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[28] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, , 2021. .

[29] Ruha Benjamin. Race after technology: Abolitionist tools for the new jim code. *Social Forces*, 2019.

[30] Glen Berman. Machine learning practices and infrastructures. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 466–481, New York, NY, USA, 2023. Association for Computing Machinery.

[31] Anthony Bernier. Representations of youth in local media: Implications for library service. *Library & Information Science Research*, 33(2):158–167, 2011.

[32] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, 56:85–100, 2014.

[33] James Bessen, Stephen Michael Impink, and Robert Seamans. The cost of ethical ai development for ai startups. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 92–106, 2022.

[34] Amy L Best. Teen driving as public drama: statistics, risk, and the social construction of youth as a public problem. *Journal of Youth Studies*, 11(6):651–669, 2008.

[35] Eshta Bhardwaj, Rohan Alexander, and Christoph Becker. Limits to ai growth: The ecological and social consequences of scaling. *arXiv preprint arXiv:2501.17980*, 2025.

[36] Sudeep Bhatia and Lukasz Walasek. Predicting implicit attitudes with natural language data. *Proceedings of the National Academy of Sciences*, 120(25):e2220726120, 2023.

[37] Ananya Bhattacharjee, Joseph Jay Williams, Jonah Meyerhoff, Harsh Kumar, Alex Mariakakis, and Rachel Kornfield. Investigating the role of context in the delivery of text messages for supporting psychological wellbeing. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[38] Justin B Biddle. On predicting recidivism: epistemic risk, tradeoffs, and values in machine learning. *Canadian Journal of Philosophy*, 52(3):321–341, 2022.

[39] Justin B Biddle and Rebecca Kukla. The geography of epistemic risk. *Exploring inductive risk: Case studies of values in science*, pages 215–237, 2017.

[40] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com.

[41] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. *URL https://openaipublic. blob. core. windows. net/neuron-explainer/paper/index. html.(Date accessed: 14.05. 2023)*, 2023.

[42] Abeba Birhane, Sepehr Dehdashtian, Vinay Prabhu, and Vishnu Boddeti. The dark side of dataset scaling: Evaluating racial classification in multimodal models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1229–1244, 2024.

[43] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. Power to the people? opportunities and challenges for participatory ai. *Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8, 2022.

[44] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184, , 2022. .

[45] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

[46] Ekaba Bisong and Ekaba Bisong. An overview of google cloud platform services. *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, ():7–10, 2019.

[47] Erling Björgvinsson, Pelle Ehn, and Per-Anders Hillgren. Participatory design and "democratizing innovation". In *Proceedings of the 11th Biennial Participatory Design Conference*, PDC '10, page 41–50, New York, NY, USA, 2010. Association for Computing Machinery.

[48] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164. Association for Computing Machinery, 2023.

[49] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[50] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357, 2016.

[51] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Koulako Bala Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel J. Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui

Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *ArXiv*, abs/2108.07258, 2021.

[52] Nadav Borenstein, Karolina Stańczak, Thea Rolskov, Natália da Silva Perez, Natacha Klein Käfer, and Isabelle Augenstein. Measuring intersectional biases in historical documents. *arXiv preprint arXiv:2305.12376*, 2023.

[53] Virginia Braun and Victoria Clarke. *Thematic analysis.* American Psychological Association, 2012.

[54] Virginia Braun and Victoria Clarke. Conceptual and design thinking for thematic analysis. *Qualitative Psychology*, 9(1):3, 2022.

[55] Virginia Braun and Victoria Clarke. Everything changes... well some things do: Reflections on, and resources for, reflexive thematic analysis. *QMiP Bulletin*, 2022.

[56] Levin Brinkmann, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, Thomas F Müller, Anne-Marie Nussberger, Agnieszka Czaplicka, Alberto Acerbi, Thomas L Griffiths, Joseph Henrich, et al. Machine culture. *Nature Human Behaviour*, 7(11):1855–1868, 2023.

[57] Stephanie Brookes and Lisa Waller. Communities of practice in the production and resourcing of fact-checking. *Journalism*, 24(9):1938–1958, 2023.

[58] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[59] Erik Brynjolfsson, Danielle Li, and Lindsey R Raymond. Generative ai at work. Technical report, National Bureau of Economic Research, 2023.

[60] Erik Brynjolfsson, Daniel Rock, and Chad Syverson. The productivity j-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, 13(1):333–372, 2021.

[61] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[62] Christy M Buchanan, Daniel Romer, Laura Wray-Lake, and Sheretta T Butler-Barnes. Adolescent storm and stress: a 21st century evaluation. *Frontiers in Psychology*, 14:1257641, 2023.

[63] Christy M Buchanan, Susannah Zietz, Jennifer E Lansford, Ann T Skinner, Laura Di Giunta, Kenneth A Dodge, Sevtap Gurdal, Qin Liu, Qian Long, Paul Oburu, et al. Typicality and trajectories of problematic and positive behaviors over adolescence in eight countries. *Frontiers in psychology*, 13:991727, 2023.

[64] Yaroslav Bulatov. Fitting larger networks into memory. `https://medium.com/tensorflow/fitting-larger-networks-into-memory-583e3c758ff9`, 2018. [Accessed 19-01-2024].

[65] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[66] Molly Burleson, Monique Martin, and Rashunda Lewis. Assessing the impact of nonviolent communication. *The Center for Nonviolent Communication [En línea] https://www. cnvc. org/sites/cnvc. org/files/NVC_ Research_ Files/[Consultado el 25 de enero de 2019]*, 2011.

[67] Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R Banaji. Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 156–170, 2022.

[68] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[69] Steve Campbell, Melanie Greenwood, Sarah Prior, Toniele Shearer, Kerrie Walkem, Sarah Young, Danielle Bywaters, and Kim Walker. Purposive sampling: complex or simple? research case examples. *Journal of research in Nursing*, 25(8):652–661, 2020.

[70] Tara Capel, Bernd Ploderer, Filip Bircanin, Simon Hanmer, Jamie Yates, Jiaxuan Wang, Kai Ling Khor, Tuck Wah Leong, Greg Wadley, and Michelle Newcomb. Studying self-care with generative ai tools: Lessons for design. *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, 2024.

[71] Stuart K Card, Jock D Mackinlay, and George G Robertson. The design space of input devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 117–124, 1990.

[72] Nicholas Carlini, Daniel Paleka, Krishnamurthy Dvijotham, Thomas Steinke, Jonathan Hayase, A. Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Eric Wallace, David Rolnick, and Florian Tramèr. Stealing part of a production language model. *ArXiv*, abs/2403.06634, 2024.

[73] Taiwan FactCheck Center. Taiwan factcheck center. `https://tfc-taiwan.org.tw/en`, 2024. [Accessed 22-01-2024].

[74] Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan.

Creativity support in the age of large language models: An empirical study involving emerging writers. *ArXiv*, abs/2309.12570, 2023.

[75] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, , 2020. .

[76] Tessa ES Charlesworth, Sa-kiera TJ Hudson, Emily J Cogsdill, Elizabeth S Spelke, and Mahzarin R Banaji. Children use targets' facial appearance to guide and predict social behavior. *Developmental Psychology*, 55(7):1400, 2019.

[77] Africa Check. Africa check. `https://africacheck.org/`, 2024. [Accessed 22-01-2024].

[78] Jerry Chee, Yaohui Cai, Volodymyr Kuleshov, and Christopher De Sa. Quip: 2-bit quantization of large language models with guarantees. *arXiv preprint arXiv:2307.13304*, ():, 2023.

[79] Mark Chen, Alec Radford, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, 2020.

[80] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[81] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, ():, 2016.

[82] Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked personas: Using natural language prompts to measure stereotypes in language models. In *Proceedings of the 61st Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1504–1532, 2023.

[83] Chequeado. Chequeado. `https://chequeado.com/`, 2024. [Accessed 22-01-2024].

[84] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*, 2022.

[85] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[86] Eun Cheol Choi and Emilio Ferrara. Automated claim matching with large language models: empowering fact-checkers in the fight against misinformation. *arXiv preprint arXiv:2310.09223*, 2023.

[87] Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. Chatgpt goes to law school. *Available at SSRN*, 2023.

[88] Amit K Chopra and Munindar P Singh. Sociotechnical systems and ethics in the large. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 48–53, 2018.

[89] Lynn Schofield Clark. *From angels to aliens: Teenagers, the media, and the supernatural*. Oxford University Press, 2005.

[90] Victoria Clarke and Virginia Braun. Thematic analysis. *The Journal of Positive Psychology*, 12:297 – 298, 2017.

[91] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12:2493–2537, 2011.

[92] ColombiaCheck. Colombiacheck. `https://colombiacheck.com/`, 2024. [Accessed 22-01-2024].

[93] Shanley Corvite, Kat Roemmich, Tillie Ilana Rosenberg, and Nazanin Andalibi. Data subjects' perspectives on emotion artificial intelligence use in the workplace: A relational ethics lens. *Proceedings of the ACM on Human-Computer Interaction*, 7:1 – 38, 2023.

[94] Paul T Costa and Robert R McCrae. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, 2(2):179–198, 2008.

[95] C̃ommon Crawl. Common crawl. `https://commoncrawl.org/`, 2024. Accessed: 2024-04-09.

[96] RMIT FactLab CrossCheck. Rmit factlab crosscheck. `https://www.rmit.edu.au/about/schools-colleges/media-and-communication/industry/factlab/crosscheck`, 2024. [Accessed 22-01-2024].

[97] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. *arXiv preprint arXiv:2204.08583*, 2022.

[98] Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. *Advances in neural information processing systems*, 28():, 2015.

[99] Aayushi Dangol, Michele Newman, Robert Wolfe, Jin Ha Lee, Julie A Kientz, Jason Yip, and Caroline Pitt. Mediating culture: Cultivating socio-cultural understanding of ai in children through participatory design. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, pages 1805–1822, 2024.

[100] Anubrata Das, Houjiang Liu, Venelin Kovatchev, and Matthew Lease. The state of human-centered nlp technology for fact-checking. *Information processing & management*, 60(2):103219, 2023.

[101] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online, June 2021. Association for Computational Linguistics.

[102] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–23, 2023.

[103] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[104] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019.

[105] Univision El Detector. Univision el detector. `https://www.univision.com/especiales/noticias/detector/index.html`, 2024. [Accessed 22-01-2024].

[106] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Llm. int8 (): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*, ():, 2022.

[107] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

[108] Tim Dettmers and Luke Zettlemoyer. The case for 4-bit precision: k-bit inference scaling laws. In *International Conference on Machine Learning*, pages 7750–7774, , 2023. PMLR, .

[109] Sunipa Dev, J. Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. Building socio-culturally inclusive stereotype resources with community engagement. *ArXiv*, abs/2307.10514, 2023.

[110] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[111] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *ArXiv*, abs/2005.00341, 2020.

[112] Laura Di Giunta, Carolina Lunetti, Jennifer E Lansford, Nancy Eisenberg, Concetta Pastorelli, Dario Bacchini, Liliana Maria Uribe Tirado, Anne-Marie R Iselin, Emanuele Basili, Giulia Gliozzo, et al. Predictors and outcomes associated with the growth curves of self-efficacy beliefs in regard to anger and sadness regulation during adolescence: a longitudinal cross-cultural study. *Frontiers in Psychology*, 14:1010358, 2023.

[113] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018.

[114] Colin Dickey. The rise and fall of facts. https://www.cjr.org/special_report/rise-and-fall-of-fact-checking.php, 2019. [Accessed 09-03-2024].

[115] Laurence Dierickx, Carl-Gustav Lindén, and Andreas Lothe Opdahl. Automated fact-checking to support professional practices: systematic literature review and meta-analysis. *International Journal of Communication*, 17:21, 2023.

[116] Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*, ():, 2020.

[117] Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235, 2023.

[118] Lori Dorfman and Vincent Schiraldi. *Off balance: Youth, race & crime in the news*. Building Blocks for Youth, 2001.

[119] Lori Dorfman, Katie Woodruff, Vivian Chavez, and Lawrence Wallack. Youth and violence on local television news in california. *American Journal of Public Health*, 87(8):1311–1316, 1997.

[120] Judith Dörrenbächer, Ronda Ringfort-Felner, and Marc Hassenzahl. The intricacies of social robots: Secondary analysis of fictional documentaries to explore the benefits and challenges of robots in complex social settings. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[121] Anil R. Doshi and Oliver P. Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10, 2024.

[122] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning*, 2017.

[123] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[124] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Bap tiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Cantón Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab A. AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriele Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guanglong Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Laurens Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Ju-Qing Jia, Kalyan Vasuden Alwala, K. Upasani, Kate Plawiak, Keqian Li, Ken-591 neth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuen ley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Babu Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew

Oldham, Mathieu Rita, Maya Pavlova, Melissa Hall Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri S. Chatterji, Olivier Duchenne, Onur cCelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasić, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Ro main Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Chandra Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whit ney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yiqian Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zhengxu Yan, Zhengxing Chen, Zoe Papakipos, Aaditya K. Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adi Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Ben Leonhardi, Po-Yao (Bernie) Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon

Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Shang-Wen Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzm'an, Frank J. Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory G. Sizov, Guangyi Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Han Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kaixing(Kai) Wu, U KamHou, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, A Lavender, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong,

Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollár, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sung-Bae Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Andrei Poenaru, Vlad T. Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xia Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models. *ArXiv*, abs/2407.21783, 2024.

[125] Mats Ekström, Amanda Ramsälv, and Oscar Westlund. The epistemologies of breaking news. *Journalism Studies*, 22(2):174–192, 2021.

[126] Mats Ekström, Amanda Ramsälv, and Oscar Westlund. Data-driven news work culture: Reconciling tensions in epistemic values and practices of news journalism. *Journalism*, 23(4):755–772, 2022.

[127] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts:

An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023.

[128] Robert D Enright, Victor M Levy, Deborah Harris, and Daniel K Lapsley. Do economic conditions influence how theorists view adolescents? *Journal of Youth and Adolescence*, 16:541–559, 1987.

[129] Ilker Etikan, Sulaiman Abubakar Musa, Rukayya Sunusi Alkassim, et al. Comparison of convenience sampling and purposive sampling. *American journal of theoretical and applied statistics*, 5(1):1–4, 2016.

[130] Factly. Factly. `https://factly.in/`, 2024. [Accessed 22-01-2024].

[131] Aos Fatos. Aos fatos. `https://www.aosfatos.org/`, 2024. [Accessed 22-01-2024].

[132] Science Feedback. Science feedback. `https://science.feedback.org/`, 2024. [Accessed 22-01-2024].

[133] Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. Generative ai. *Business & Information Systems Engineering*, 66(1):111–126, 2024.

[134] Robert G Fichman and Chris F Kemerer. The illusory diffusion of innovation: An examination of assimilation gaps. *Information systems research*, 10(3):255–275, 1999.

[135] Code for Africa. Code for africa. `https://github.com/CodeForAfrica/`, 2024. [Accessed 22-01-2024].

[136] The Center for Nonviolent Communication. The center for nonviolent communication. `https://www.cnvc.org/`, 2024. [Accessed 02-08-2024].

[137] Valentina Franzoni and Alfredo Milani. Emotion recognition for self-aid in addiction treatment, psychotherapy, and nonviolent communication. In *Computational Science and Its Applications–ICCSA 2019: 19th International Conference, Saint Petersburg, Russia, July 1–4, 2019, Proceedings, Part II 19*, pages 391–404. Springer, 2019.

[138] Batya Friedman, David G. Hendry, and Alan Borning. A survey of value sensitive design methods. *Found. Trends Hum.-Comput. Interact.*, 11(2):63–125, nov 2017.

[139] Yue Fu, Sami Foell, Xuhai Xu, and Alexis Hiniker. From text to self: Users' perceptions of potential of ai on interpersonal communication and self. *arXiv preprint arXiv:2310.03976*, 2023.

[140] Yue Fu, Mingrui Ray Zhang, Lynn K. Nguyen, Yifan Lin, Rebecca Michelson, Tala June Tayebi, and Alexis Hiniker. Self-talk with superhero zip: Supporting children's socioemotional learning with conversational agents. *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*, 2023.

[141] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[142] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[143] Lise Getoor and Ashwin Machanavajjhala. Entity resolution: theory, practice & open challenges. *Proc. VLDB Endow.*, 5(12):2018–2019, aug 2012.

[144] Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feedforward layers build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680*, 2022.

[145] Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscope: A unifying framework for inspecting hidden representations of language models. *arXiv preprint arXiv:2401.06102*, 2024.

[146] Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy. *Advances in neural information processing systems*, 34:27131–27145, 2021.

[147] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023.

[148] Barry Glassner. *The culture of fear: Why Americans are afraid of the wrong things: Crime, drugs, minorities, teen moms, killer kids, muta.* Hachette UK, 2010.

[149] Maoguo Gong, Yu Xie, Ke Pan, Kaiyuan Feng, and Alex Kai Qin. A survey on differentially private machine learning. *IEEE computational intelligence magazine*, 15(2):49–64, 2020.

[150] Google. Google + team usa — dear sydney. `https://www.youtube.com/watch?v=NgtHJKn0Mck`, 2024. [Accessed 31-07-2024].

[151] John R Graham, Campbell R Harvey, and Manju Puri. A corporate beauty contest. *Management Science*, 63(9):3044–3056, 2017.

[152] Grammarly. Grammarly. `https://www.grammarly.com/`, 2024. [Accessed 22-01-2024].

[153] Édouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomáš Mikolov. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[154] Lucas Graves. Anatomy of a fact check: Objective practice and the contested epistemology of fact checking. *Communication, culture & critique*, 10(3):518–537, 2017.

[155] Lucas Graves. Factsheet: Understanding the promise and limits of automated fact-checking. *Reuters Inst. Study of Journalism, Univ. Oxford, Oxford*, 2018.

[156] Lucas Graves and Michelle A. Amazeen. Fact-checking as idea and practice in journalism. *Oxford Research Encyclopedia of Communication*, 2019.

[157] GROK. Grok the world. `https://groktheworld.com/`, 2024. [Accessed 02-08-2024].

[158] Lu Guan, Weiying Shi, Qianqian Li, Jeffry Oktavianus, and Mengmeng Wu. Have color representations in books changed over the past 200 years? an empirical analysis based on the google books ngram corpus. *Color Research & Application*, 49(1):65–78, 2024.

[159] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[160] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022.

[161] Carlos Ignacio Gutierrez, Anthony Aguirre, Risto Uuk, Claire C. Boine, and Matija Franklin. A proposal for a definition of general purpose artificial intelligence systems. *Digital Society*, 2, 2023.

[162] Yuval Haber, Inbar Levkovich, Dorit Hadar-Shoval, and Zohar Elyoseph. The artificial third: a broad view of the effects of introducing generative artificial intelligence on psychotherapy. *JMIR Mental Health*, 11:e54781, 2024.

[163] Bin Han, Haotian Zhu, Sitong Zhou, Sofia Ahmed, Md Mushfiqur Rahman, Fei Xia, and Kevin Lybarger. Huskyscribe at mediqa-sum 2023: Summarizing clinical dialogues with transformers. In *Conference and Labs of the Evaluation Forum*, 2023.

[164] LynNell Hancock. The school shootings: Why context counts. *Columbia Journalism Review*, 40(1):76–76, 2001.

[165] Kenneth R Hanson and Hannah Bolthouse. "replika removing erotic role-play is like grand theft auto removing guns or cars": Reddit discourse on artificial intelligence chatbots and sexual technologies. *Socius*, 10:23780231241259627, 2024.

[166] Hossein Hassani and Emmanuel Sirmal Silva. The role of chatgpt in data science: how ai-assisted conversational interfaces are revolutionizing the field. *Big data and cognitive computing*, 7(2):62, 2023.

[167] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[168] Will Douglas Heaven. Why meta's latest large language model survived only three days online. `https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/`, 2022. [Accessed 12-07-2024].

[169] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

[170] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *ArXiv*, abs/2104.08718, 2021.

[171] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

[172] Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, 2015.

[173] Kashmir Hill. Openai worries about what its chatbot will say about people's faces. *The New York Times*, Jul 2023.

[174] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531, 2015.

[175] Michael Jeffrey Daniel Hoefer and Stephen Voida. Being, having, doing, and interacting: A personal informatics approach to understanding human need satisfaction in everyday life. *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 2023.

[176] Linda Holmes. The antithesis of the olympics: Using ai to write a fan letter. `https://www.npr.org/2024/07/30/nx-s1-5056201/google-olympics-ai-ad`, 2024. [Accessed 31-07-2024].

[177] Ari Holtzman, Peter West, and Luke Zettlemoyer. Generative models as a complex systems science: How can we make sense of large language model behavior? *arXiv preprint arXiv:2308.00189*, 2023.

[178] Aspen Hopkins and Serena Booth. Machine learning practices outside big tech: How resource constraints challenge responsible development. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, page 134–145, New York, NY, USA, 2021. Association for Computing Machinery.

[179] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander J. Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *ArXiv*, abs/2305.02301, 2023.

[180] Long-Jing Hsu, Janice K Bays, Katherine M Tsui, and Selma Sabanovic. Co-designing social robots with people living with dementia: Fostering identity, connectedness, security, and autonomy. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, pages 2672–2688, 2023.

[181] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang,

Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, ():, 2021.

[182] Jane Im, Jill Dimond, Melody Berton, Una Lee, Katherine Mustelier, Mark S Ackerman, and Eric Gilbert. Yes: Affirmative consent as a theoretical framework for understanding and imagining social platforms. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–18, 2021.

[183] Infoveritas. Infoveritas. `https://info-veritas.com/`, 2024. [Accessed 22-01-2024].

[184] Daphne Ippolito, Reno Kriz, João Sedoc, Maria Kustikova, and Chris Callison-Burch. Comparison of diverse decoding methods from conditional language models. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762, Florence, Italy, July 2019. Association for Computational Linguistics.

[185] Sebastian Jäckle, Thomas Metz, Georg Wenzelburger, and Pascal D König. A catwalk to congress? appearance-based effects in the elections to the us house of representatives 2016. *American Politics Research*, 48(4):427–441, 2020.

[186] Bastian Jaeger, Alexander T Todorov, Anthony M Evans, and Ilja van Beest. Can we reduce facial biases? persistent effects of facial trustworthiness on sentencing decisions. *Journal of Experimental Social Psychology*, 90:104004, 2020.

[187] Shrey Jain, Connor Spelliscy, Samuel Vance-Law, and Scott Moore. Ai and democracy's digital identity crisis. *arXiv preprint arXiv:2311.16115*, 2023.

[188] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–15, 2023.

[189] Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. Personalized

soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*, 2023.

[190] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55:1 – 38, 2022.

[191] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, ():, 2023.

[192] Harry H Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. Ai art and its impact on artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 363–374, 2023.

[193] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

[194] Brian D Johnson and Ryan D King. Facial profiling: Race, physical appearance, and punishment. *Criminology*, 55(3):520–547, 2017.

[195] Prerna Juneja and Tanushree Mitra. Human and technological infrastructures of fact-checking. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–36, 2022.

[196] Kimberly Barsamian Kahn, Phillip Atiba Goff, J Katherine Lee, and Diane Motamed. Protecting whiteness: White phenotypic racial stereotypicality reduces police use of force. *Social Psychological and Personality Science*, 7(5):403–411, 2016.

[197] Ruth Kansky and Tarek Maassarani. Teaching nonviolent communication to increase empathy between people and toward wildlife to promote human–wildlife coexistence. *Conservation Letters*, 15(1):e12862, 2022.

[198] S Kapoor and A Narayanan. How to prepare for the deluge of generative ai on social media, 2023.

[199] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[200] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020.

[201] Ryan Kelly, Daniel Gooch, Bhagyashree Patil, and Leon Watts. Demanding by design: Supporting effortful communication practices in close personal relationships. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 70–83, 2017.

[202] Eugenia Kim, De'Aira Bryant, Deepak Srikanth, and Ayanna Howard. Age bias in emotion detection: An analysis of facial emotion recognition performance on young, middle-aged, and older adults. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 638–644, 2021.

[203] JaeWon Kim, Soobin Cho, Robert Wolfe, Jishnu Hari Nair, and Alexis Hiniker. Privacy as social norm: Systematically reducing dysfunctional privacy concerns on social media. *arXiv preprint arXiv:2410.16137*, 2024.

[204] JaeWon Kim, Robert Wolfe, Ishita Chordia, Katie Davis, and Alexis Hiniker. " sharing, not showing off": How bereal approaches authentic self-presentation on social media through its design. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–32, 2024.

[205] Rebecca Klar. Teens use, hear of chatgpt more than parents: poll, 2023.

[206] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. Chatgpt: Jack of all trades, master of none. *Information Fusion*, ():101861, 2023.

[207] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664, , 2021. PMLR, .

[208] Pravesh Koirala and Nobal B. Niraula. NPVec1: Word embeddings for Nepali - construction and evaluation. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 174–184, Online, August 2021. Association for Computational Linguistics.

[209] Philippe Laban, Lidiya Murakhovs' ka, Caiming Xiong, and Chien-Sheng Wu. Are you sure? challenging llms leads to performance drops in the flipflop experiment. *arXiv preprint arXiv:2311.08596*, 2023.

[210] Rabindra Lamsal. 300-dimensional word embeddings for nepali language. *IEEE Dataport*, 2019.

[211] Max Langenkamp and Daniel N Yue. How open source machine learning software shapes ai. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 385–395, 2022.

[212] Reed Larson and Suzanne Wilson. Adolescence across place and time. *Handbook of adolescent psychology*, pages 297–330, 2004.

[213] Patrick Yung Kang Lee, Ning F. Ma, Ig-Jae Kim, and Dongwook Yoon. Speculating on risks of ai clones to selfhood and relationships: Doppelganger-phobia, identity

fragmentation, and living memories. *Proceedings of the ACM on Human-Computer Interaction*, 7:1 – 28, 2023.

[214] Gabriel S Lenz and Chappell Lawson. Looking the part: Television leads less informed citizens to vote based on candidates' appearance. *American Journal of Political Science*, 55(3):574–589, 2011.

[215] Simon A Levin, Helen V Milner, and Charles Perrings. The dynamics of political polarization, 2021.

[216] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[217] Han Li and Renwen Zhang. Finding love in algorithms: deciphering the emotional contexts of close encounters with ai chatbots. *Journal of Computer-Mediated Communication*, 29(5):zmae015, 2024.

[218] Hanlin Li, Nicholas Vincent, Stevie Chancellor, and Brent Hecht. The dimensions of data labor: A road map for researchers, activists, and policymakers to empower data producers. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1151–1161, New York, NY, USA, 2023. Association for Computing Machinery.

[219] Jiexun Li and Xiaohui Chang. Combating misinformation by sharing the truth: a study on the spread of fact-checks on social media. *Information systems frontiers*, 25(4):1479–1493, 2023.

[220] Jingjin Li, Nayeon Kwon, Huong Pham, Ryun Shim, and Gilly Leshed. Co-designing magic machines for everyday mindfulness with practitioners. *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 2023.

[221] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[222] Lan Li, Tina Lassiter, Joohee Oh, and Min Kyung Lee. Algorithmic hiring in practice: Recruiter and hr professional's perspectives on ai use in hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 166–176, 2021.

[223] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models, 2023.

[224] Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. Opening up chatgpt: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th international conference on conversational user interfaces*, pages 1–6, 2023.

[225] Stephanie C. Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Annual Meeting of the Association for Computational Linguistics*, 2021.

[226] Ellinor Lindqvist, Eva Pettersson, and Joakim Nivre. To the most gracious highness, from your humble servant: Analysing swedish 18th century petitions using text classifica-

tion. In *Proceedings of the 6th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 53–64, 2022.

[227] Litmus. Litmus. `https://litmus-factcheck.jp/about/en/`, 2024. [Accessed 22-01-2024].

[228] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[229] logically.ai. logically.ai. `https://www.logically.ai/`, 2024. [Accessed 22-01-2024].

[230] Chiara Longoni, Andrey Fradkin, Luca Cian, and Gordon Pennycook. News from generative artificial intelligence is believed less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 97–106, 2022.

[231] S. Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, 2023.

[232] Lefteris Loukas, Ilias Stogiannidis, Prodromos Malakasiotis, and Stavros Vassos. Breaking the bank with ChatGPT: Few-shot text classification for finance. In Chung-Chi Chen, Hiroya Takamura, Puneet Mathur, Remit Sawhney, Hen-Hsen Huang, and Hsin-Hsi Chen, editors, *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 74–80, Macao, 20 August 2023. -.

[233] Nicola Lucchi. Chatgpt: a case study on copyright challenges for generative artificial intelligence systems. *European Journal of Risk Regulation*, 15(3):602–624, 2024.

[234] Kai Lukoff, Ulrik Lyngs, and Lize Alberts. Designing to support autonomy and reduce psychological reactance in digital self-control tools. In *Position Papers for the Workshop*

"Self-Determination Theory in HCI: Shaping a Research Agenda" at the Conference on Human Factors in Computing Systems (CHI'22), volume 5, 2022.

[235] Kai Lukoff, Ulrik Lyngs, Himanshu Zade, J Vera Liao, James Choi, Kaiyue Fan, Sean A Munson, and Alexis Hiniker. How the design of youtube influences user sense of agency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2021.

[236] Maldita. Maldita. `https://maldita.es/`, 2024. [Accessed 22-01-2024].

[237] Mike A Males. *Framing youth: Ten myths about the next generation.* Common Courage Press, 1999.

[238] Gabriela Marcu and Jina Huh-Yoo. Attachment-informed design: Digital interventions that build self-worth, relationships, and community in support of mental health. *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 2023.

[239] Elizabeth Marlow, Adeline Nyamathi, William T Grajeda, Newt Bailey, Amanda Weber, and Jerry Younger. Nonviolent communication training and empathy in male parolees. *Journal of Correctional Health Care*, 18(1):8–19, 2012.

[240] Alice E Marwick. To catch a predator? the myspace moral panic. *First Monday*, 2008.

[241] Meedan. Meedan. `https://meedan.com/`, 2024. [Accessed 22-01-2024].

[242] Shahan Ali Memon and Jevin D West. Search engines post-chatgpt: How generative artificial intelligence could make search less reliable. *arXiv preprint arXiv:2402.11707*, 2024.

[243] Lisa Messeri and MJ Crockett. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002):49–58, 2024.

[244] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[245] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

[246] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[247] Jessica K. Miller, Batya Friedman, and Gavin Jancke. Value tensions in design: the value sensitive design, development, and appropriation of a corporation's groupware system. *Proceedings of the 2007 ACM International Conference on Supporting Group Work*, 2007.

[248] Joseph Millum and Michael Garnett. How payment for research participation can be coercive. *The American Journal of Bioethics*, 19(9):21–31, 2019.

[249] MindaNews. Mindanews. `https://www.mindanews.com/`, 2024. [Accessed 22-01-2024].

[250] Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *ArXiv*, abs/2310.17884, 2023.

[251] Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 174–184, 2018.

[252] Kirsten Morehouse, Vaibhav Rouduri, Wil Cunningham, and Tessa Charlesworth. Traces of human attitudes in contemporary and historical word embeddings (1800-2000). *Research Square Preprint*, 2023.

[253] Christopher P Morley and Sam Grammer. Now more than ever: reflections on the state and importance of peer review. *PRiMER: Peer-review reports in medical education research*, 5, 2021.

[254] Meredith Ringel Morris. Scientists' perspectives on the potential for generative ai in their fields. *arXiv preprint arXiv:2304.01420*, 2023.

[255] Meredith Ringel Morris, Carrie J Cai, Jess Holbrook, Chinmay Kulkarni, and Michael Terry. The design space of generative models. *arXiv preprint arXiv:2304.10547*, 2023.

[256] Megan Morrone. Google olympics ad sparks new ire over generative ai. `https://www.axios.com/2024/07/31/google-olympics-ad-ai-gemini-ire`, 2024. [Accessed 31-07-2024].

[257] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

[258] Michael Muller, Melanie Feinberg, Timothy George, Steven J Jackson, Bonnie E John, Mary Beth Kery, and Samir Passi. Human-centered study of data science work practices. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*, pages 1–8, 2019.

[259] Anne-Claire Museux, Serge Dumont, Emmanuelle Careau, and Élise Milot. Improving interprofessional collaboration: The effect of training in nonviolent communication. *Social work in health care*, 55(6):427–439, 2016.

[260] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

[261] Mahin Naderifar, Hamideh Goli, and Fereshteh Ghaljaie. Snowball sampling: A purposeful method of sampling in qualitative research. *Strides in development of medical education*, 14(3), 2017.

[262] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Ouyang Long, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. Webgpt: Browser-assisted question-answering with human feedback. *ArXiv*, abs/2112.09332, 2021.

[263] Terrence Neumann and Nicholas Wolczynski. Does ai-assisted fact-checking disproportionately benefit majority groups online? In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 480–490, 2023.

[264] Newtral. Newtral. `https://www.newtral.es/`, 2024. [Accessed 22-01-2024].

[265] US NIH. Nih style guide: Age. `https://www.nih.gov/nih-style-guide/age`, 2022. Accessed: 2023-10-01.

[266] Marcianna Nosek, Elizabeth Gifford, and B. Kober. Nonviolent communication (nvc) training increases empathy in baccalaureate nursing students: A mixed method study. *Journal of Nursing Education and Practice*, 4:1, 2014.

[267] Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023.

[268] Caroline Mimbs Nyce. Google wins the gold medal for worst olympic ad. `https://www.theatlantic.com/technology/archive/2024/07/google-dear-sydney-olympic-ad/679292/`, 2024. [Accessed 31-07-2024].

[269] Fabian Offert and Thao Phan. A sign that spells: Dall-e 2, invisual images and the racial politics of feature space. *arXiv preprint arXiv:2211.06323*, 2022.

[270] Office of the U.S. Surgeon General. Our epidemic of loneliness and isolation. https://www.hhs.gov/sites/default/files/surgeon-general-social-connection-advisory.pdf, 2023. [Accessed 02-08-2024].

[271] DongWon Oh, Elinor A Buck, and Alexander Todorov. Revealing hidden gender biases in competence impressions of faces. *Psychological Science*, 30(1):65–79, 2019.

[272] David B Olawade, Ojima Z Wada, Aderonke Odetayo, Aanuoluwapo Clement David-Olawade, Fiyinfoluwa Asaolu, and Judith Eberhardt. Enhancing mental health with artificial intelligence: Current trends and future prospects. *Journal of medicine, surgery, and public health*, page 100099, 2024.

[273] Christopher Y Olivola and Alexander Todorov. Elected in 100 milliseconds: Appearance-based trait inferences and voting. *Journal of nonverbal behavior*, 34:83–110, 2010.

[274] Etienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. Chatgpt for text annotation? mind the hype! *SocArXiv. October*, 4, 2023.

[275] Étienne Ollion, Rubing Shen, Ana Macanovic, and Arnault Chatelain. The dangers of using proprietary llms for research. *Nature Machine Intelligence*, pages 1–2, 2024.

[276] Jesutofunmi A. Omiye, Jenna C. Lester, Simon Spichak, Veronica M Rotemberg, and Roxana Daneshjou. Large language models propagate race-based medicine. *NPJ Digital Medicine*, 6, 2023.

[277] Shiva Omrani Sabbaghi and Aylin Caliskan. Measuring gender bias in word embeddings of gendered languages requires disentangling grammatical gender signals. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 518–531, 2022.

[278] Shiva Omrani Sabbaghi, Robert Wolfe, and Aylin Caliskan. Evaluating biased attitude associations of language models in an intersectional context. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 542–553, 2023.

[279] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[280] OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin,

Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2023.

[281] OpenAI. Introducing chatgpt. *OpenAI Blog*, ():, Nov 2022.

[282] OpenAI. Reducing bias and improving safety in dall·e 2, 2022.

[283] OpenAI. Fine-tuning. `https://platform.openai.com/docs/guides/fine-tuning`, 2024. [Accessed 13-09-2024].

[284] OpenAI. Gpt-4o system card, 2024.

[285] OpenAI. Introducing the gpt store. *OpenAI Blog*, ():, Jan 2024.

[286] OpenAI. Learning to reason with llms. `https://openai.com/index/learning-to-reason-with-llms/`, 2024. [Accessed 12-09-2024].

[287] OpenAI. Memory and new controls for chatgpt. *OpenAI Blog*, ():, Feb 2024.

[288] OpenAI. Models. `https://platform.openai.com/docs/models/`, 2024. [Accessed 19-01-2024].

[289] OpenAI. Pricing. `https://openai.com/pricing`, 2024. [Accessed 19-01-2024].

[290] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[291] Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? *ArXiv*, abs/2309.05196, 2023.

[292] Rachel Pain. Youth, age and the representation of fear. *Capital & class*, 27(2):151–171, 2003.

[293] Alexis Palmer, Noah A Smith, and Arthur Spirling. Using proprietary language models in academic research requires explicit justification. *Nature Computational Science*, 4(1):2–3, 2024.

[294] Orestis Papakyriakopoulos, Christelle Tessono, Arvind Narayanan, and Mihir Kshirsagar. How algorithms shape the distribution of political advertising: Case studies of facebook, google, and tiktok. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 532–546, 2022.

[295] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, , 2002. .

[296] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*, 2021.

[297] Joon Sung Park, Michael S Bernstein, Robin N Brewer, Ece Kamar, and Meredith Ringel Morris. Understanding the representation and representativeness of age in ai data sets. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 834–842, 2021.

[298] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

[299] Sohyun Park, Anja Thieme, Jeongyun Han, Sungwoo Lee, Wonjong Rhee, and Bongwon Suh. "i wrote as if i were telling a story to someone i knew.": Designing chatbot interactions for expressive writing in mental health. *Proceedings of the 2021 ACM Designing Interactive Systems Conference*, 2021.

[300] Dylan Patel and Gerald Wong. Gpt-4 architecture, infrastructure, training dataset, costs, vision, moe. `https://www.semianalysis.com/p/gpt-4-architecture-infrastructure`, 2023. [Accessed 19-01-2024].

[301] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, ():, 2023.

[302] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[303] Charith Peris, Christophe Dupuy, Jimit Majmudar, Rahil Parikh, Sami Smaili, Richard Zemel, and Rahul Gupta. Privacy in the time of language models. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1291–1292, , 2023. .

[304] Joshua C Peterson, Stefan Uddenberg, Thomas L Griffiths, Alexander Todorov, and Jordan W Suchow. Deep models of superficial face judgments. *Proceedings of the National Academy of Sciences*, 119(17):e2115228119, 2022.

[305] Angela Phillips. Transparency and the new ethics of journalism. *Journalism Practice*, 4(3):373–382, 2010.

[306] Dustin Podell, Zion English, Kyle Lacey, A. Blattmann, Tim Dockhorn, Jonas Muller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *ArXiv*, abs/2307.01952, 2023.

[307] Pagella Politica. Pagella politica. `https://pagellapolitica.it/`, 2024. [Accessed 22-01-2024].

[308] Politifact. Politifact. `https://www.politifact.com/`, 2024. [Accessed 22-01-2024].

[309] Jon Porter. Chatgpt continues to be one of the fastest-growing services ever. *The Verge*, ():, Nov 2023.

[310] Poynter. Commit to transparency — sign up for the international fact-checking network's code of principles. https://ifcncodeofprinciples.poynter.org, 2024. [Accessed 09-03-2024].

[311] Poynter. International fact checking network. https://www.poynter.org/ifcn/, 2024. [Accessed 22-01-2024].

[312] Poynter. Verified signatories of the ifcn code of principles. https://ifcncodeofprinciples.poynter.org/signatories, 2024. [Accessed 22-01-2024].

[313] Kalyan Prasad Agrawal. Towards adoption of generative ai in organizational settings. *Journal of Computer Information Systems*, pages 1–16, 2023.

[314] Pravda. Pravda. https://pravda.org.pl/, 2024. [Accessed 22-01-2024].

[315] Australian Associated Press. Australian associated press. https://www.aap.com.au/, 2024. [Accessed 22-01-2024].

[316] Agence France Presse. Agence france presse. https://www.afp.com/en, 2024. [Accessed 22-01-2024].

[317] Brian A Primack, Sabrina A Karim, Ariel Shensa, Nicholas Bowman, Jennifer Knight, and Jaime E Sidani. Positive and negative experiences on social media and perceived social isolation. *American Journal of Health Promotion*, 33(6):859–868, 2019.

[318] Yang Qu, Eva M Pomerantz, Qian Wang, and Florrie Fei-Yin Ng. Early adolescents' stereotypes about teens in hong kong and chongqing: Reciprocal pathways with problem behavior. *Developmental psychology*, 56(6):1092, 2020.

[319] Organizers Of Queerinai, Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubicka, Hang Yuan, Hetvi J, Huan Zhang, Jaidev Shriram, Kruno Lehman, Luca Soldaini, Maarten Sap, Marc Peter Deisenroth, Maria Leonor Pacheco, Maria Ryskina, Martin Mundt,

Milind Agarwal, Nyx Mclean, Pan Xu, A Pranav, Raj Korpan, Ruchira Ray, Sarah Mathew, Sarthak Arora, St John, Tanvi Anand, Vishakha Agrawal, William Agnew, Yanan Long, Zijie J. Wang, Zeerak Talat, Avijit Ghosh, Nathaniel Dennler, Michael Noseworthy, Sharvani Jha, Emi Baylor, Aditya Joshi, Natalia Y. Bilenko, Andrew Mcnamara, Raphael Gontijo-Lopes, Alex Markham, Evyn Dong, Jackie Kay, Manu Saraswat, Nikhil Vytla, and Luke Stark. Queer in ai: A case study in community-led participatory ai. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1882–1895, New York, NY, USA, 2023. Association for Computing Machinery.

[320] The Quint. The quint. `https://www.thequint.com/international`, 2024. [Accessed 22-01-2024].

[321] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[322] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

[323] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. , ():, 2018.

[324] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[325] Evani Radiya-Dixit and Gina Neff. A sociotechnical audit: Assessing police use of facial recognition. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1334–1346, 2023.

[326] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[327] Chahat Raj, Anjishnu Mukherjee, and Ziwei Zhu. True and fair: Robust and unbiased fake news detection via interpretable machine learning. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 962–963, 2023.

[328] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151, 2020.

[329] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[330] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[331] Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. Fairness in language models beyond english: Gaps and challenges. *arXiv preprint arXiv:2302.12578*, 2023.

[332] Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37, 2023.

[333] Rappler. Rappler. `https://www.rappler.com/`, 2024. [Accessed 22-01-2024].

[334] Mary Elizabeth Raven and Alicia Flanders. Using contextual inquiry to learn about your audiences. *ACM SIGDOC Asterisk Journal of Computer Documentation*, 20(1):1–13, 1996.

[335] Partha Pratim Ray. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, ():, 2023.

[336] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[337] Thomson Reuters. Thomson reuters. `https://www.thomsonreuters.com/en.html`, 2024. [Accessed 22-01-2024].

[338] Rev. Rev. `https://www.rev.com/`, 2024. [Accessed 10-03-2024].

[339] Jenn Riley. Understanding metadata. *Washington DC, United States: National Information Standards Organization (http://www. niso. org/publications/press/UnderstandingMetadata. pdf)*, 23:7–10, 2017.

[340] Paavo Ritala, Mika Ruokonen, and Laavanya Ramaul. Transforming boundaries: how does chatgpt change knowledge work? *Journal of Business Strategy*, 2023.

[341] Adi Robertson. 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`, 2024. [Accessed 12-07-2024].

[342] Anna Rogers, Niranjan Balasubramanian, Leon Derczynski, Jesse Dodge, Alexander Koller, Sasha Luccioni, Maarten Sap, Roy Schwartz, Noah A. Smith, and Emma Strubell. Closed ai models make bad baselines, Apr 2023.

[343] Carl R Rogers. The concept of the fully functioning person. *Psychotherapy: Theory, Research & Practice*, 1(1):17, 1963.

[344] Carl R Rogers. Toward a science of the person. *Journal of Humanistic Psychology*, 3(2):72–92, 1963.

[345] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[346] Marshall B Rosenberg and Deepak Chopra. *Nonviolent communication: A language of life: Life-changing tools for healthy relationships*. PuddleDancer Press, 2015.

[347] Marshall B Rosenberg and Riane Eisler. *Life-enriching education: Nonviolent communication helps schools improve performance, reduce conflict, and enhance relationships*. PuddleDancer Press, 2003.

[348] Annabel Rothschild, Amanda Meng, Carl Disalvo, Britney Johnson, Ben Rydal Shapiro, and Betsy Disalvo. Interrogating data work as a community of practice. *Proceedings of the ACM on Human-Computer Interaction*, 6:1 – 28, 2022.

[349] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[350] Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264, 2023.

[351] Sayed Saniya Salim, Agrawal Nidhi Ghanshyam, Darkunde Mayur Ashok, Dungarpur Burhanuddin Mazahir, and Bhushan S Thakare. Deep lstm-rnn with word embedding for sarcasm detection on twitter. In *2020 international conference for emerging technology (INCET)*, pages 1–4. IEEE, 2020.

[352] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey,

M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng-Xin Yong, Harshit Pandey, Michael McKenna, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, page , , 2022. .

[353] Takanori Sano and Hideaki Kawabata. A computational approach to investigating facial attractiveness factors using geometric morphometric analysis and deep learning. *Scientific reports*, 13(1):19797, 2023.

[354] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.

[355] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[356] Brennan Schaffner, Yaretzi Ulloa, Riya Sahni, Jiatong Li, Ava Kim Cohen, Natasha Messier, Lan Gao, and Marshini Chetty. An experimental study of netflix use and the effects of autoplay on watching behaviors. *arXiv preprint arXiv:2412.16040*, 2024.

[357] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551, 2023.

[358] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wight-

man, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[359] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[360] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*, 2023.

[361] Hanieh Shakeri, Ye Yuan, Benett Axtell, Denise Y Geiskkovitch, and Carman Neustaedter. Designing smart home technology for passive co-presence over distance. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, pages 3389–3406, 2024.

[362] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.

[363] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, et al. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2023.

[364] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, 2019.

[365] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Scharli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, 2023.

[366] Donghoon Shin, Soomin Kim, Ruoxi Shang, Joonhwan Lee, and Gary Hsieh. Introbot: Exploring the use of chatbot-assisted familiarization in online collaborative groups. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[367] Snopes. Snopes. `https://www.snopes.com/`, 2024. [Accessed 09-03-2024].

[368] Irene Solaiman. The gradient of generative ai release: Methods and considerations. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, pages 111–122, 2023.

[369] Der Spiegel. Der spiegel. `https://www.spiegel.de/international/`, 2024. [Accessed 22-01-2024].

[370] Clay Spinuzzi. The methodology of participatory design. *Technical Communication*, 52:163–174, 05 2005.

[371] Laurence Steinberg, Grace Icenogle, Elizabeth P Shulman, Kaitlyn Breiner, Jason Chein, Dario Bacchini, Lei Chang, Nandita Chaudhary, Laura Di Giunta, Kenneth A Dodge, et al. Around the world, adolescence is a time of heightened sensation seeking and immature self-regulation. *Developmental science*, 21(2):e12532, 2018.

[372] Susannah R Stern. Self-absorbed, dangerous, and disengaged: What popular films tell us about teenagers. *Mass Communication & Society*, 8(1):23–38, 2005.

[373] Susannah R Stern and Sarah Burke Odland. Constructing dysfunction: News coverage of teenagers and social media. *Mass Communication and society*, 20(4):505–525, 2017.

[374] Janka I Stoker, Harry Garretsen, and Luuk J Spreeuwers. The facial appearance of ceos: Faces signal selection but not performance. *PloS one*, 11(7):e0159950, 2016.

[375] Lead Stories. Lead stories. `https://leadstories.com/`, 2024. [Accessed 22-01-2024].

[376] Alejandra Suarez, Dug Y Lee, Christopher Rowe, Alex Anthony Gomez, Elise Murowchick, and Patricia L Linn. Freedom project: Nonviolent communication and mindfulness training in prison. *Sage Open*, 4(1):2158244013516154, 2014.

[377] Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache, 2019.

[378] Bipesh Subedi and Prakash Poudyal. Word embedding in nepali language using word2vec. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '22, page 152–156, New York, NY, USA, 2023. Association for Computing Machinery.

[379] Manal Sultan, Lia Jacobs, Abby Stylianou, and Robert Pless. Exploring clip for real world, text-based image retrieval. *2023 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–6, 2023.

[380] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[381] Clare AM Sutherland and Andrew W Young. Understanding trait impressions from faces. *British Journal of Psychology*, 113(4):1056–1078, 2022.

[382] Brian W Swider, T Brad Harris, and Qing Gong. First impression effects in organizational psychology. *Journal of Applied Psychology*, 2021.

[383] Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 305–311, 2019.

[384] Felicia Fang-Yi Tan, Ashwin Ram, Chloe Haigh, and Shengdong Zhao. Mindful moments: Exploring on-the-go mindfulness practice on smart-glasses. *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 2023.

[385] Charlotte Tang, Yunan Chen, Bryan C Semaan, and Jahmeilah A Roberson. Restructuring human infrastructure: The impact of ehr deployment in a volunteer-dependent clinic. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 649–661, 2015.

[386] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony S. Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *ArXiv*, abs/2211.09085, 2022.

[387] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[388] Tech4Peace. Tech4peace. `https://t4p.co/`, 2024. [Accessed 22-01-2024].

[389] Eva H Telzer, Junqiang Dai, Jimmy J Capella, Maria Sobrino, and Shedrick L Garrett. Challenging stereotypes of teens: Reframing adolescence as window of opportunity. *American Psychologist*, 77(9):1067, 2022.

[390] Doris Hooi Ten Wong, Chen Siang Phang, Nurazean Maarop, Ganthan Narayana Samy, Roslina Ibrahim, Rasimah Che Mohd Yusoff, Pritheega Magalingam, and Nurulhuda Firdaus Mohd Azmi. Effect of social media on human interpersonal communication: A review. *Open International Journal of Informatics (OIJI)*, 5(2):1–6, 2017.

[391] Rosamond Thalken, Edward Stiglitz, David Mimno, and Matthew Wilkens. Modeling legal reasoning: Lm annotation at the edge of human agreement. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9252–9265, , 2023. .

[392] The Guardian. Ron desantis signs florida social media ban for children into law, 2024.

[393] The New York Times. The chatgpt lawyer explains himself. `https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html`, 2023. [Accessed 22-01-2024].

[394] Sulav Timilsina, Milan Gautam, and Binod Bhattarai. Nepberta: Nepali language model trained in a large corpus. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 273–284, 2022.

[395] India Today. India today. `https://www.indiatoday.in/`, 2024. [Accessed 22-01-2024].

[396] Alexander Todorov. Face value. In *Face Value*. Princeton University Press, 2017.

[397] Connor Toups, Rishi Bommasani, Kathleen A. Creel, Sarah H. Bana, Dan Jurafsky, and Percy Liang. Ecosystem-level analysis of deployed machine learning reveals homogeneous outcomes. *ArXiv*, abs/2307.05862, 2023.

[398] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, ():, 2023.

[399] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, ():, 2023.

[400] Andreas Tsamados, Nikita Aggarwal, Josh Cowls, Jessica Morley, Huw Roberts, Mariarosaria Taddeo, and Luciano Floridi. The ethics of algorithms: key problems and solutions. *Ethics, governance, and policies in artificial intelligence*, pages 97–123, 2021.

[401] Frank Tsui, Orlando Karam, and Barbara Bernal. *Essentials of software engineering*. Jones & Bartlett Learning, 2022.

[402] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*, ():, 2023.

[403] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[404] April Tyack and Elisa D Mekler. Self-determination theory in hci games research: Current uses and open questions. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–22, 2020.

[405] Anja Van den Broeck, D Lance Ferris, Chu-Hsiang Chang, and Christopher C Rosen. A review of self-determination theory's basic psychological needs at work. *Journal of management*, 42(5):1195–1229, 2016.

[406] Maarten Vansteenkiste, Richard M Ryan, and Bart Soenens. Basic psychological need theory: Advancements, critical themes, and future directions, 2020.

[407] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[408] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Will we run out of data? limits of llm scaling based on human-generated data. *arXiv preprint arXiv:2211.04325*, 3, 2024.

[409] Emily Vogels and Risa Gelles-Watnick. Teens and social media: Key findings from pew research center surveys, 2023.

[410] Carl Vondrick. Assisting iclr 2025 reviewers with feedback. https://blog.iclr.cc/2024/10/09/iclr2025-assisting-reviewers/, 2024. [Accessed 06-11-2024].

[411] Renata Wacker and Isabel Dziobek. Preventing empathic distress and social stressors at work through nonviolent communication training: A field study with health professionals. *Journal of occupational health psychology*, 23(1):141, 2018.

[412] Wen wai Yim, Asma Ben Abacha, Griffin Adams, Neal Snider, and Meliha Yetisgen-Yildiz. Overview of the mediqa-sum task at imageclef 2023: Summarization and classification of doctor-patient conversations. In *Conference and Labs of the Evaluation Forum*, 2023.

[413] Ethan Waisberg, Joshua Ong, Nasif Zaman, Sharif Amit Kamran, Prithul Sarker, Alireza Tavakkoli, and Andrew G Lee. Gpt-4 for triaging ophthalmic symptoms. *Eye*, 37(18):3874–3875, 2023.

[414] Nathan Walter, Jonathan Cohen, R Lance Holbert, and Yasmin Morag. Fact-checking: A meta-analysis of what works and for whom. *Political communication*, 37(3):350–375, 2020.

[415] Dakuo Wang, Justin D Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–24, 2019.

[416] Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI Conference on Artificial Intelligence*, 2022.

[417] Tianlu Wang, Xi Victoria Lin, Nazneen Rajani, Vicente Ordonez, and Caimng Xiong. Double-hard debias: Tailoring word embeddings for gender bias mitigation. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

[418] Xi Wang, Laurence Aitchison, and Maja Rudolph. Lora ensembles for large language model fine-tuning. *ArXiv*, abs/2310.00035, 2023.

[419] Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*, ():, 2023.

[420] Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*, ():, 2021.

[421] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652, 2021.

[422] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[423] Jerry W. Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. Simple synthetic data reduces sycophancy in large language models. *ArXiv*, abs/2308.03958, 2023.

[424] Laura Weidinger, John F. J. Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zachary Kenton, Sande Minnich Brown, William T. Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models. *ArXiv*, abs/2112.04359, 2021.

[425] Justin D Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer. Design principles for generative ai applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–22, 2024.

[426] Bingbing Wen, Bill Howe, and Lucy Lu Wang. Characterizing llm abstention behavior in science qa with context perturbations. *arXiv preprint arXiv:2404.12452*, 2024.

[427] Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. The art of refusal: A survey of abstention in large language models. *arXiv preprint arXiv:2407.18418*, 2024.

[428] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France, May 2020. European Language Resources Association.

[429] WhatsApp. Whatsapp. `https://www.whatsapp.com/`, 2024. [Accessed 10-03-2024].

[430] Janine Willis and Alexander Todorov. First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological science*, 17(7):592–598, 2006.

[431] John Paul Wilson and Nicholas O Rule. Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological science*, 26(8):1325–1331, 2015.

[432] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, ():, 2019.

[433] Robert Wolfe, Mahzarin R. Banaji, and Aylin Caliskan. Evidence for hypodescent in visual semantic ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

[434] Robert Wolfe and Aylin Caliskan. American==white in multimodal language-and-image ai. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022.

[435] Robert Wolfe and Aylin Caliskan. Contrastive visual semantic pretraining magnifies the semantics of natural language representations. *Association for Computational Linguistics*, 2022.

[436] Robert Wolfe and Aylin Caliskan. Detecting emerging associations and behaviors with regional and diachronic word embeddings. In *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*, pages 91–98. IEEE, 2022.

[437] Robert Wolfe and Aylin Caliskan. Markedness in visual semantic ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022.

[438] Robert Wolfe and Aylin Caliskan. Vast: The valence-assessing semantics test for contextualizing language models. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, 2022.

[439] Robert Wolfe, Aayushi Dangol, Alexis Hiniker, and Bill Howe. Dataset scale and societal consistency mediate facial impression bias in vision-language ai. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, 2024.

[440] Robert Wolfe, Aayushi Dangol, Bill Howe, and Alexis Hiniker. Representation bias of adolescents in ai: A bilingual, bicultural study. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, 2024.

[441] Robert Wolfe and Alexis Hiniker. Expertise fog on the gpt store: Deceptive design patterns in user-facing generative ai. *Mobilizing Research and Regulatory Action on*

Dark Patterns and Deceptive Design Practices Workshop at CHI Conference on Human Factors in Computing Systems, 2024.

[442] Robert Wolfe, Alexis Hiniker, and Bill Howe. Ml-eat: A multilevel embedding association test for interpretable and transparent social science. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, 2024.

[443] Robert Wolfe and Tanushree Mitra. The impact and opportunities of generative ai in fact-checking. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1531–1543, 2024.

[444] Robert Wolfe and Tanushree Mitra. The implications of open generative models in human-centered data science work: A case study with fact-checking organizations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1595–1607, 2024.

[445] Robert Wolfe, Isaac Slaughter, Bin Han, Bingbing Wen, Yiwei Yang, Lucas Rosenblatt, Bernease Herman, Eva Brown, Zening Qu, Nic Weber, et al. Laboratory-scale ai: Open-weight models are competitive with chatgpt even in low-resource settings. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1199–1210, 2024.

[446] Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1174–1185, 2023.

[447] Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Daniel Jurafsky, Christopher D. Manning, and Christopher Potts. Reft: Representation finetuning for language models. *ArXiv*, abs/2404.03592, 2024.

[448] Lukas Wutschitz, Huseyin A. Inan, and Andre Manoel. dp-transformers: Training transformer models with differential privacy. https://www.microsoft.com/en-us/research/project/dp-transformers, August 2022.

[449] Yu Xia, Tong Yu, Zhankui He, Handong Zhao, Julian McAuley, and Shuai Li. Aligning as debiasing: Causality-aware alignment via reinforcement learning with interventional feedback. In *North American Chapter of the Association for Computational Linguistics*, 2024.

[450] Sally Y Xie, Jessica K Flake, Ryan M Stolier, Jonathan B Freeman, and Eric Hehman. Facial impressions are predicted by the structure of group stereotypes. *Psychological Science*, 32(12):1979–1993, 2021.

[451] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.

[452] Anna Xygkou, Panote Siriaraya, Alexandra Covaci, Holly Gwen Prigerson, Robert A. Neimeyer, Chee Siang Ang, and Wan Jou She. The "conversation" about loss: Understanding how chatbot technology was used in supporting people in grief. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[453] Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. Physiognomy in the age of ai. *Feminist AI: Critical Perspectives on Algorithms, Data, and Intelligent Machines*, page 208, 2023.

[454] Yiwei Yang, Anthony Z Liu, Robert Wolfe, Aylin Caliskan, and Bill Howe. Label-efficient group robustness via out-of-distribution concept curation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12426–12434, 2024.

[455] Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. Alignment for honesty. *ArXiv*, abs/2312.07000, 2023.

[456] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, et al. Opacus: User-friendly differential privacy library in pytorch. *arXiv preprint arXiv:2109.12298*, ():, 2021.

[457] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, ():, 2021.

[458] Xinying Yu, Shi Xu, and Mark Ashton. Antecedents and outcomes of artificial intelligence adoption and application in the workplace: the socio-technical system theory perspective. *Information Technology & People*, 36(1):454–474, 2023.

[459] Yangyang Yu and Jordan Suchow. Deep tensor factorization models of first impressions. In *SVRHM 2022 Workshop@ NeurIPS*, 2022.

[460] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *ArXiv*, abs/2311.16502, 2023.

[461] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022.

[462] G Zagni and T Canetta. Generative ai marks the beginning of a new era for disinformation, 2023.

[463] Hubert D Zajkac, Dana Li, Xiang Dai, Jonathan F Carlsen, Finn Kensing, and Tariq O Andersen. Clinician-facing ai in the wild: Taking stock of the sociotechnical challenges and opportunities for hci. *ACM Transactions on Computer-Human Interaction*, 30(2):1–39, 2023.

[464] Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, , 2022. .

[465] Giacomo Zanotti, Daniele Chiffi, and Viola Schiaffonati. Ai-related risk: An epistemological approach. *Philosophy & Technology*, 2024.

[466] Antonia Zapf, Stefanie Castell, Lars Morawietz, and André Karch. Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology*, 16(1):93, August 2016.

[467] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *ArXiv*, abs/1409.2329, 2014.

[468] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022.

[469] Xiao Zhan, Yifan Xu, and Ştefan Sarkadi. Deceptive ai ecosystems: The case of chatgpt. *Proceedings of the 5th International Conference on Conversational User Interfaces*, 2023.

[470] Alex Wuqi Zhang, Ting-Han Lin, Xuan Zhao, and Sarah Strohkorb Sebo. Ice-breaking technology: Robots and computers can foster meaningful connections between strangers through in-person conversations. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[471] Amy X. Zhang, Michael Muller, and Dakuo Wang. How do data science workers collaborate? roles, workflows, and tools. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1), may 2020.

[472] Renwen Zhang, Kathryn E. Ringland, Melina Paan, David C. Mohr, and Madhu Reddy. Designing for emotional well-being: integrating persuasion and customization into mental health technologies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

[473] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, page , , 2019. .

[474] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

[475] Zhiping Zhang, Michelle Jia, Hao-Ping Lee, Bingsheng Yao, Sauvik Das, Ada Lerner, Dakuo Wang, and Tianshi Li. "it's a fair game", or is it? examining how users navigate disclosure risks and benefits when using llm-based conversational agents. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2023.

[476] Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online, July 2020. Association for Computational Linguistics.

[477] Tony Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, 2021.

[478] Wei Zhao, Ryan M Kelly, Melissa J Rogerson, and Jenny Waycott. Older adults using technology for meaningful activities during covid-19: An analysis through the lens of self-determination theory. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2023.

[479] Yunpu Zhao, Rui Zhang, Junbin Xiao, Changxin Ke, Ruibo Hou, Yifan Hao, Qi Guo, and Yunji Chen. Towards analyzing and mitigating sycophancy in large vision-language models. *ArXiv*, abs/2408.11261, 2024.

[480] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric. P Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. Lmsys-chat-1m: A large-scale real-world llm conversation dataset, 2023.

[481] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

[482] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18697–18709, 2022.

[483] Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore, December 2023. Association for Computational Linguistics.

[484] Zoom. Zoom. `https://zoom.us/`, 2024. [Accessed 10-03-2024].