Imitation, Identity, and Injustice in Artificial Intelligence

by

Jackie Kay

Submitted to the Department of Computer Science
on February 26th, 2025 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Thesis supervisor: Marc Deisenroth

Title: Professor of Computer Science

Thesis supervisor: Shakir Mohamed

Title: Director for Science, Technology & Society, Google Deepmind

ABSTRACT

Replicating human behavior is a popular goal of AI system design. Imitation learning is an established subfield dedicated to this objective, in which a neural network is optimized to imitate trajectories from a data distribution originating from an expert. Human-like qualities can also emerge unexpectedly, due to properties of the training data or other design parameters. This unintentional imitation–or the failure to achieve the goal of imitation—can have undesirable consequences when AI is deployed in the real world.

This thesis explores the imitation of humans in artificial intelligence, from its technical dimensions to philosophical questions and ethical implications. To first illustrate imitation as a method, I present research on building a general-purpose motor intelligence: a dataset gathered by teleoperation, and a foundation model trained via imitation learning. I then widen my concerns to imitation as a goal, by studying how machines might imitate human social identity. Many existing classification systems fail to operationalize a nuanced theory of identity, resulting in the exacerbation of social injustice. I propose technical interventions for meeting our proposed definition of identity. Finally, I turn away from aiming to imitate human qualities, instead studying how injustice typically perpetrated by humans emerges in AI. I draw on the established philosophical theory of epistemic injustice to study how unique forms of it arise in applications of generative AI. I conclude by imagining what the field of artificial intelligence could look like beyond the anthropomorphic boundaries of imitation.

Throughout this thesis, I alternate between perspectives on imitation as a method, goal, and an emergence to gain a holistic insight on how the automated imitation of humanity impacts all levels of society–whether intentional or accidental. This triangulation enables an interdisciplinary reflection on the technical and ethical responsibilities of machine learning practitioners.

I, Jackie Kay, confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Impact Statement

The research presented in this thesis has the potential to contribute to the development of more human-compatible and socially responsible AI systems. Within academia, the critical analysis of identity representation in AI and the exploration of epistemic injustice in generative AI can stimulate further research in AI fairness and ethics, encouraging the development of systems that are more inclusive, transparent, and accountable.

Beyond academia, the thesis has implications for the development and deployment of AI systems in various domains. The insights into the ethical challenges of AI imitation can inform policymakers, industry leaders, and technologists in shaping responsible AI practices. By recognizing the potential for AI to perpetuate social injustices and epistemic harms, stakeholders can proactively work towards mitigating these risks and ensuring that AI technologies are used to promote equity and fairness. The thesis also encourages a broader societal dialogue about the role of AI in our lives, emphasizing the importance of critical engagement and informed decision-making in navigating the complex landscape of AI and its impact on human identity and knowledge. In conclusion, this thesis challenges the field of AI to move beyond mere imitation and towards a more nuanced understanding of human behavior and cognition. It calls for a critical and ethical approach to AI development, emphasizing the importance of inclusivity, transparency, and accountability in building AI systems that benefit society as a whole.

Acknowledgments

First of all, I thank my advisors: Marc Deisenroth, who supported me through the whole process, Raia Hadsell, who enabled me to start the program, and Shakir Mohamed, who anchored me through the last year of the PhD after mentoring me unofficially for years. I am hugely grateful to all of my co-authors, especially my dear friends Mel Vecerik and Christina Lu, and my colleague and mentor Atoosa Kasirzadeh. I also acknowledge the PhD programme team at Google DeepMind for funding and supporting the administrative side my PhD.

My upgrade viva committee Lourdes Agapito and Dimitrios Kanoulis provided valuable mid-point feedback. I additionally thank Cynthia Matusek and Jack Stilgoe for serving on my viva committee.

The Sustainable/Statistical Machine Learning group provided a wonderful academic and collegial environment, especially Maria Perez-Ortiz, Dan Giles, and the other students in my cohort, Yicheng Luo and Sicelukwanda Zwane. I am also thankful for professional support from my mentors at DeepMind, including Matt Hoffman, Gabe Barth-Maron, and my line manager Tom Stepleton, as well as Scott Reed, Nando de Freitas, and Konrad Zolna for their excellent leadership and allowing me to include excerpts from Gato in this thesis.

Obviously, none of this would have been possible without the love and support of my parents, and my sibling, who inspired me to think philosophically from a young age. I especially want to thank my late grandmother Myung-ok Pearl Kay and my late grandfather Bill Hughes. Knowing how proud it would have made them to get the first doctorate on both sides of our family helped me achieve everything I have today. And finally, thank you to my partner, Carolina, for always letting me know it was going to be okay–whether I finished this thing or not!

UCL Research Paper Declaration Form: A Generalist Agent

1. **For a research manuscript that has already been published:**

   (a) **What is the title of the manuscript?** A Generalist Agent

   (b) **Please include a link to or doi for the work:** https://openreview.net/forum?id=1ikK0kHjvj

   (c) **Where was the work published?** Transactions on Machine Learning Research (TMLR)

   (d) **Who published the work?** OpenReview

   (e) **When was the work published?** 10th November, 2022

   (f) **List the manuscript's authors in the order they appear on the publication:** Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, Nando de Freitas

   (g) **Was the work peer reviewed?** Yes

   (h) **Have you retained the copyright?** Yes

   (i) **Was an earlier form of the manuscript uploaded to a preprint serve?** https://arxiv.org/abs/2205.06175

2. **For multi-authored work, please give a statement of contribution covering all authors:** A full statement of contribution can be found within the original published manuscript [1].

3. **In which chapter(s) of your thesis can this material be found?** Chapter 3

UCL Research Paper Declaration Form: Subverting Machines, Fluctuating Identities

1. **For a research manuscript that has already been published :**

   (a) **What is the title of the manuscript?** Subverting machines, fluctuating identities: Re-learning human categorization

   (b) **Please include a link to or doi for the work:** https://dl.acm.org/doi/abs/10.1145/3531146.3533161

   (c) **Where was the work published?** ACM Conference on Fairness, Accountability, and Transparency (FAccT)

   (d) **Who published the work?** Association for Computing Machinery (ACM)

   (e) **When was the work published?** 20th June 2022

   (f) **List the manuscript's authors in the order they appear on the publication:** Christina Lu, Jackie Kay, Kevin McKee

   (g) **Was the work peer reviewed?** Yes

   (h) **Have you retained the copyright?** Yes

   (i) **Was an earlier form of the manuscript uploaded to a preprint server?** https://arxiv.org/abs/2205.13740

2. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:** Jackie Kay contributed the sections "The Limitations of AI Fairness" and "Alternative System Configurations", and co-wrote the introduction, conclusion, and abstract with Christina Lu. Christina Lu wrote the section "Identity as Autopoiesis" and "Theorizing about Identity". Kevin McKee co-wrote "Technical Approaches to Closing the Loop" with Jackie Kay.

3. **In which chapter(s) of your thesis can this material be found?** Chapter 4

UCL Research Paper Declaration Form: Epistemic Injustice in Generative AI

1. **1. For a research manuscript that has already been published:**

   (a) **What is the title of the manuscript?** Epistemic Injustice in Generative AI

   (b) **Please include a link to or doi for the work:** https://ojs.aaai.org/index.php/AIES/article/view/31671

   (c) **Where was the work published?** AAAI/ACM Conference on AI, Ethics, and Society

   (d) **Who published the work?** Association for the Advancement of Artificial Intelligence (AAAI)

   (e) **When was the work published?** 16th October 2024

   (f) **List the manuscript's authors in the order they appear on the publication:** Jackie Kay, Atoosa Kasirzadeh, Shakir Mohamed

   (g) **Was the work peer reviewed?** Yes

   (h) **Have you retained the copyright?** No

   (i) **Was an earlier form of the manuscript uploaded to a preprint server?** https://arxiv.org/abs/2408.11441

2. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)**:** Jackie Kay researched the theory and examples, originated the key theoretical ideas, and authored the majority of the text. Atoosa Kasirzadeh contributed the taxonomy framing of the paper and originated the subtype of access injustice. Shakir Mohamed was senior supervisor, providing valuable feedback and oversight.

3. **In which chapter(s) of your thesis can this material be found?** Chapter 5

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Imitation as goal, method, and emergence

The imitation of human behavior and cognition is a major theme in the field of artificial intelligence. In 1950, Alan Turing proposed his "imitation game" (later called the Turing Test) as a thought experiment for answering the question, "can machines think?" [2]. The proposal for the Dartmouth Summer Research Project on Artificial Intelligence, a founding event in the field, suggests that machines can be made to "simulate" the human brain [3]. Machine imitation of humans continues to play a critical role in the field today, although the term has deepened in meaning from Turing's original usage.

The field of imitation learning offers straightforward tools for learning to reproduce human behavior. Starting from datasets that record actions and their surrounding context, techniques such as behavioral cloning are designed to optimize a policy to imitate these demonstrations [4]. Thus imitation can be viewed as a learning method, an algorithmic strategy for acquiring established skills. However, imitation learning has limitations: it cannot be used to discover entirely new strategies for accomplishing tasks, and is fundamentally limited by the quality of the demonstrations – if demonstrations are suboptimal for a given task, the imitator will not learn the optimal strategy [5]. Furthermore, these algorithms may overfit by exploiting properties of the data that are not representative of the desired behavior [5]. Imitation is also fundamentally difficult when the embodiment and circumstances of a robot or computational agent differ from that of a human, due to the mismatch of morphology

and control dimensionality [6].

While imitation of a data distribution is the goal of imitation learning, I also distinguish the case where imitation is the goal of the system, but not the explicit optimization rule. That is, in imitation learning, replicating the action-observation correlations from the training dataset is the exact goal of developing the system. However, in imitation as a goal, the overarching goal of developing the system is the reproduction of some specific quality which may be hard to describe analytically or even record in a dataset. In the case of imitation-as-goal, imitation is the design target. It can also be an evaluative measure: how much does the system's behavior resemble certain desirable human traits? Using alternative methods to imitation learning to may help to resolve the overfitting issues described above. However, human-imitating technology often leaves users in an uncanny valley [7], or misled to think the technology's inner workings resemble a human, when only some observable component or capability does. The anthropomorphic design of chatbots and other natural language systems leads to the obfuscation of the system's inner workings, and a misunderstanding of the system as a conscious, cognizing agent [8].

Human imitation in AI system is not always intentional. While an imitator may accidentally diverge from the goal of cloning human behavior, there may also be accidentally convergence. Thus imitation may be an emergent property of the system, arising from implicit factors in its design, development and deployment. According to Perrow's normal accident theory, accidental emergence of unexpected behaviors is an inevitable in complex systems [9]. Systems that are not designed to mimic humans will reflect the values of their creators, and often enact consequences that replicate the preferences of humans [10]. The embedding of human values and behaviors into a technology is obvious, not unexpected, with the realization that society interacts with the technological system at every step of the way.

The question of whether or not machines *should* imitate humans is charged with ethical considerations. The impact of a technology can be characterized as positive or negative (but in reality, these consequences do not fit into strict and objective binaries). Experts have predicted that artificial intelligence will have a far-reaching impact on all aspects of our society [11]. Even if the creators of this technology are motivated by a desire for positive impact, good intentions are not enough to prevent the misuse or unexpected consequences of

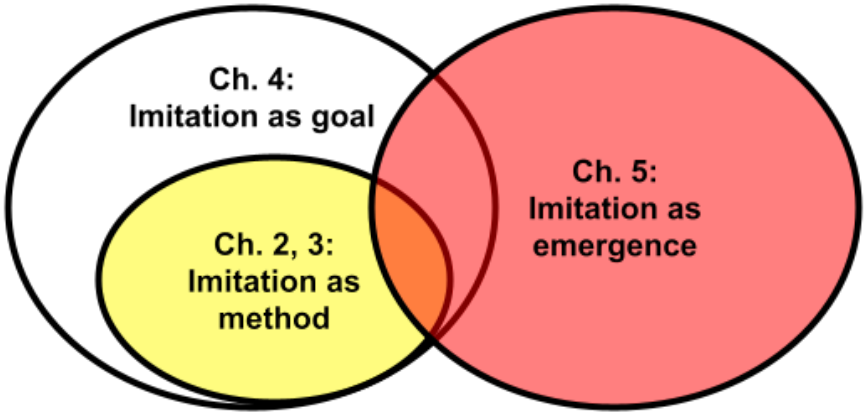

Figure 1.1: Overlapping relationships between the three distinct conceptualizations of imitation as explored in the corresponding chapters of this thesis: method, goal, and emergence.

a new technology – take as an example the amplification and worsening of systemic racial bias through algorithmic classification and decision-making [12]. The unconscious biases, cultural assumptions, and moral values of human designers have a tendency to creep into the system they create, exacerbating social injustices via technology [13]. Therefore, technologists have a responsibility to introspect on these factors and incorporate an interdisciplinary ethical practice into the research process. Given the possibility for emergence, it also implies an ethical commitment to evaluating and monitoring systems for unexpected imitation.

Of course, the ethical commitments of technologists far extend beyond evaluation. These considerations must be embedded into the design, implementation, and usages of a technology. The tense geopolitics of AI have fomented growing alliances between technology companies and global superpowers, who seek to consolidate power by incorporating AI into national security, surveillance, and military uses. Thus it is critical to maintain a practice of reflection and direct action against a potential wave of violent applications. To that end, my research into imitation will question the basic premise of why we build this technology and explore ethical dimensions of representation and justice.

The technical considerations and ethical impacts of imitation motivate a multi-faceted approach to design solutions and recommendations. Three different definitions of "imitation" – as a method, goal, and emergence – capture the intersecting concerns of the project of reproducing human behavior in AI. Figure 1.1 illustrates the intersecting relationships between these different meanings and locates the chapters of this thesis within. While imitation is

always the goal of the method of imitation learning, imitation as a method can be inscribed strictly within imitation as a goal. However, imitation of certain characteristics can emerge as an unintentional side effect of system development. As I will show, these three distinct meanings of imitation enable a holistic examination of the each stage of the AI development process, from dataset gathering and training to evaluation and wider impact analysis.

1.1.1 Thesis Outline

In this thesis, I explore these three distinct conceptualizations of the same word – imitation as method, goal, and emergence – weaving technical challenges with ethical and philosophical considerations. To illustrate imitation as a method, I present technical efforts towards building a general-purpose motor intelligence: a dataset gathered by teleoperation (Chapter 2), and a large-scale foundation model trained via imitation learning (Chapter 3). I then widen my concerns to imitation as a goal, by studying how machines might imitate human social identity. Chapter 4 analyzes how many existing classification systems fail to operationalize a nuanced conception of identity, and proposes technical interventions for meeting my proposed definition. Finally, in Chapter 5, I study how the undesired imitation of human qualities can lead to injustices perpetrated by AI. Specifically, I draw on the established philosophical theory of epistemic injustice and study how unique forms of algorithmic epistemic injustice arise in applications of generative AI, from large language models to image generation. Chapter 6 reflects the philosophical and ethical findings of the second half of the thesis back on the technical artifacts described in the first half, and exploring the wider implications of my research given recent military conflicts and political developments that threaten human rights and expressions of diverse identities. I conclude with a call to action for concerned technologists to radically shift their design practices in order to help remake the world.

1.1.2 Anthropomorphic Robot Imitation

The first two chapters of this thesis focus on steps towards developing a robotics system that exhibits general dexterity on par with human motor skills: in short, a robot that can manipulate anything a human can.

Inspired by how humans and animals acquire motor skills, imitation learning is my method of choice for the proposed system. Humans are among the most imitative creatures on the planet, and learning from observing others is one of our earliest strategies for skill acquisition [14]. Unlike most robots today, humans (and many animals) exhibit the ability to imitate others with bodies unlike their own: people of different shapes and sizes, children imitating adults, or even imitation of different species or inanimate objects [15]. In the field of robotics, learning from demonstrations and imitation learning are common strategies for replicating the behavior of experts, which could include human teleoperators, handcrafted controllers, previously learned policies, or even different embodiments [16].

Chapter 2 describes the starting point for learning general-purpose manipulation from imitation: gathering a dataset. A five-finger teleoperation device was used to control a simulated anthropomorphic robotic hand/arm. Seven different teleoperators manipulated a diverse set of 25 objects for a total of four hours of recorded demonstrations. Experiments were conducted to analyze the dataset's characteristics, including intrinsic dimensionality and hand-object dynamics. These results showed that a high diversity of grasping strategies was necessary for manipulating the wide variety of objects. Data of this kind could be used to train imitation policies for an anthropomorphic robotic hand. Anthropomorphic robot hand datasets are scarce in the field, and this dataset sets itself apart by the range of objects and motor skills it offers.

1.1.3 Generalist Imitation Agent

While the third-person teleoperation method used for this dataset produces high-quality data which is immediately useful for the robot, the process also does not scale to the amount of data that would be required for high-dimensional, dexterous, general-purpose control of a robot hand, because it is time-consuming and requires skilled demonstrators to operate an expensive and specialized apparatus. What about techniques that can leverage data collected in different contexts, with different embodiments? Foundation models trained on huge amounts of data from the internet represent a paradigm shift for few-shot and in-context learning, especially in the text and image domains [17]. What if we could apply similar techniques to sensorimotor control of a robot? This would involve a training an action

sequence model on a massive volume of "in-the-wild" videos in such a way that would enable cross-domain generalization.

Chapter 2 presents Gato, a cross-embodiment foundation model for actions. It is trained to imitate the actions of experts in a dataset of 63 million episodes from 596 separate tasks. Using similar techniques as recent advances in large-scale language modelling, Gato is 1.2 billion parameter policy network which operates on multiple modalities (images, text, proprioceptive observations, and actions). It decides based on its observed context which action space to activate, and successfully completes tasks in different environments such as games, simulations, and a real-world robotics block stacking scenario. Gato can quickly adapt to new scenarios and tasks with a small amount of data (10–100 episodes).

Does Gato actually learn useful representations that transfer across environments, or does the network simply use its large capacity to memorize effective strategies? Are the similarity relationships between different domains reflected in Gato's emergent representations? These questions are investigated using visualizations drawn from gathering the attention maps and intermediate activations of Gato. This analysis is a basic first step towards an evaluation of AI that relies on the inner workings of the model, rather than the external observed behavior.

While scale and improved architectures have enabled large-scale generalist agents, we must anticipate the ethical implications of generalist AI deployment. Robots with general-purpose manipulation skills are potentially beneficial to society, since they enable automation of a broad class of tasks that may be dangerous, repetitive, or cannot be fulfilled due to labor shortages [18]. However, the safe deployment of embodied general-purpose AI presents many challenging technical, social, and ethical problems. If agents like Gato are deployed in production scenarios, they will inevitably interact with people. In its current incarnation, Gato is trained offline on a massive dataset that may contain significant biases. Therefore, the actions of this system could be influenced by biases that encode stereotypes, unjust associations, or erasure of minoritized and marginalized identities [19]. The next part of the thesis explores the representation of social identity in AI and why it is a matter of ethical concern.

1.1.4 Representing Social Identity

A sociotechnical approach to machine learning research must acknowledge the enmeshed nature of technology and society and their circular influence on a technological artifact [20]. By consulting areas such as critical theory, sociology, and philosophy, we can broaden our understanding of machine learning system beyond the rigidly drawn parameters of quantitative science. Human biases are embedded throughout technology, which in turn influences and interacts with society, often reproducing hierarchical and colonial paradigms at massive, global scale [21]. Through a better characterization of this relationship, we can predict the ethical implications of its release and widespread adoption, and intervene on any potential negative outcomes with novel technical designs, policy recommendations, and new narratives for informing and empowering the public [22].

A sociotechnical practice is particularly urgent in an era of swift technological change due to artificial intelligence. A pressing concern with AI is algorithmic bias, a known phenomenon where data-driven algorithms reproduce systemic bias in the training data, reflecting and perpetuating social harms and injustice [23]. The field of AI fairness seeks to counteract this phenomenon by balancing outcomes for members of protected and/or marginalized groups. However, as I will show, AI fairness has fundamental limitations when it comes to representing the very identities it attempts to protect.

Chapter 4 explores how a nuanced conception of human social identity expands far beyond the representational power of current AI systems. Critical theory, gender studies, and queer theory acknowledge how identity is socially constructed through iterative discourse between a self and others. This process implies that identity can be fluid, changing between social contexts, existing above and between distinct pigeonholes of identity groups. However, identity in machine learning is often represented as a singular discrete, static property; an essential trait of the described entity. This insufficiency in the design of identity systems risks the under-representation and erasure of individuals who do not fall into the strict categories reified by AI. A critical survey of the AI fairness literature reveals the discrete, static, and essential character of how identity markers are represented at the system level.

This work proposes alternative system designs that could better describe the richness of

human social identity. The framework of multilevel optimization is explored as discursive interplay which reflects the performative dialogue between identity and the identified, and systems which make use of multiple objectives are explored as candidate algorithms for identity expression. I also propose a dataset for relational learning of identity, where a rich collection of different perspectives on identity, including the self-identification of data subjects, is collected in freeform natural language responses. While this research predated much of the recent advances in large language modelling, popular hype around today's generative models might hope that their natural language understanding would be able to parse the expansive and dynamic nuances of human social identity. However, much work remains on the ethical issues of bias and power imbalance in generative AI.

1.1.5 Epistemic Injustice in Generative AI

While Chapter 4 theorized about how AI could imitate identity representations, such imitation of human social dynamics is not always desirable or intentional. It would be unethical to reproduce the injustices that humans inflict upon others on a regular basis. However, similar patterns of prejudice held by humans emerge in the technologies we develop, often along identity lines.

Chapter 5 investigates the emergent imitation of human prejudice in generative AI. Combined with the potential for AI to spread misinformation and damage our collective knowledge, the human biases imbued in AI can result in disproportionate injustice done to already marginalized groups. The main philosophical framework behind this analysis is epistemic injustice, which examines the ethical harms done by unfairly obstructing access to knowledge, either by disbelieving the testimonies of marginalized groups, or through the misrepresentation and misunderstanding of their experiences. With this philosophical grounding, I introduce an account of *generative algorithmic epistemic injustice*: epistemic injustice inflicted by generative AI systems.

This work advocates for epistemic *justice* in generative AI, proposing strategies for resistance, system design principles, and two approaches for employing generative AI to foster a more equitable information ecosystem. The remediation of algorithmic epistemic injustice, and harms of AI more broadly, is not necessarily to engineer a "better" imitation of humanity,

but to consider a broader set of ethical principles and empower the most marginalized users of technology.

1.2 Publications included in this thesis

- Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar and Nando de Freitas. 2022. A generalist agent. In Proceedings of Transactions on Machine Learning (TMLR '22), 2835-8856. I was a secondary contributor to this project, and have included the parts of the research which I was responsible for: evaluating Gato on simulated and real robotics tasks, attention and embedding visualizations, and writing the Broader Impact section. The model description in Section 3.2 and the literature review in Section 3.4 was not originally written by myself, but I have edited them and included in this thesis for the purposes of completeness.

- Jackie Kay, Christina Lu, and Kevin McKee. 2022. Subverting machines, fluctuating identities: Re-learning human categorization. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 1005–1015. I was joint first author of this work, contributing the sections "The Limitations of AI Fairness" and "Alternative System Configurations", and co-writing the introduction, conclusion, and abstract. I have also re-written the introduction and augmented the conclusion for the purposes of this thesis.

- Jackie Kay, Atoosa Kasirzadeh, and Shakir Mohamed, 2024. Epistemic Injustice in Generative AI. In Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES '24). Association for Computing Machinery, New York, NY, USA. I was the first author of this work. I researched the theory and examples, originated the key theoretical ideas, and authored the majority of the text.

1.2.1 Publications not included in this thesis

- Nenad Tomasev, Kevin R. McKee, Jackie Kay, and Shakir Mohamed. 2021. Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21). Association for Computing Machinery, New York, NY, USA, 254–265.

- Mel Vecerik, Jackie Kay, Raia Hadsell, Lourdes Agapito, and Jon Scholz. Few-Shot Keypoint Detection as Task Adaptation via Latent Embeddings. In 2022 International Conference on Robotics and Automation (ICRA), pp. 1251-1257. IEEE, 2022.

- Anaelia Ovalle, Arjun Subramonian, Ashwin Singh, Claas Voelcker, Danica J. Sutherland, Davide Locatelli, Eva Breznik, Filip Klubička, 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 Dǒng, Jackie Kay, Manu Saraswat, Nikhil Vytla, and Luke Stark. 2023. Queer In AI: A Case Study in Community-Led Participatory AI. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 1882-1895. 2023.

- Yicheng Luo, Jackie Kay, Edward Grefenstette, and Marc Peter Deisenroth. Finetuning from Offline Reinforcement Learning: Challenges, Trade-offs and Practical Solutions. arXiv preprint arXiv:2303.17396 (2023).

- Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Connor Griffin, Ben Bariach, and Iason Gabriel. Sociotechnical Safety Evaluation of Generative AI Systems. arXiv preprint arXiv:2310.11986 (2023).

- Daniel Van Niekerk, Maria Peréz Ortiz, John Shaw-Taylor, Davor Orlic, Ivana Drobnjak, Jackie Kay, Noah Siegel, Katherine Evans, Nyalleng Moorosi, Tina Eliassi-Rad, Leone Maria Tanczer, Wayne Holmes, Marc Peter Deisenroth, Isabel Straw, Maria Fasli, Rachel Adams, Nuria Oliver, Dunja Mladenić, Urvashi Aneja. Challenging Systematic Prejudices: An Investigation into Bias Against Women and Girls. UNESCO, IRCAI (2024). Paris, France.

1.2.2 Usage of Generative AI Writing Tools

I acknowledge the use of Google's Gemini system (https://gemini.google.com/) in this thesis. Gemini was used to generate initial drafts which I then fact-checked and edited for content and style. This was done only for the Impact Statement, the Introduction of Chapter 2, and for Chapter 6.

Chapter 2

A Demonstration Dataset for General Anthropomorphic Manipulation

The precursor to imitation learning is a dataset of demonstrations: trajectories that show the desired behavior to imitate. While human demonstrations can enable effective control via data-driven solutions such as imitation learning, existing datasets for complex robot hands have numerous limitations. We introduce a simulation dataset featuring an anthropomorphic robotic hand attached to a 6-DOF arm controlled via teleoperation – the remote control of a robot via device. The dataset contains vision, touch, and proprioceptive observations for 4 hours of demonstrations from 7 unique demonstrators who manipulate a diverse set of 25 objects according to two task types: achieving target object poses and stacking. I analyze how the intrinsic dimensionality of the dataset varies with respect to objects. I also demonstrate that the dataset can be used for training dynamics models and analyze how different observations affect prediction error.

2.1 Introduction

The first step towards implementing imitation learning is a high quality dataset that demonstrates the desired behavior. In this chapter, I begin my exploration of imitation as a method by addressing a fundamental challenge in robotics: controlling a complex robotic hand to dexterously manipulate objects.

While humans perform many dexterous object manipulation tasks with ease, the control of high-dimensional artificial hands remains an open challenge in robotics. The human hand is kinematically complex, with 5 fingers, 27 degrees of freedom (DOFs), and underactuation due to joint coupling. Demonstrating fine motor skills on a robot hand requires a sophisticated level of cognition, and achieving human-level motor skills is a motivating problem for robotics. A convenient way to compare robotic object manipulation capabilities to human motor skills is to use a robot with a similar embodiment to a human (referred to as an anthropomorphic hand). Such designs overcome the limitations of one-dimensional parallel robotic grippers, which may be simpler and less costly to make, but highly constrain manual dexterity. Additionally, an anthropomorphic robot hand is a natural platform for gathering demonstrations with teleoperation. A teleoperation method which preserves the action dimensions of the human hand allows for the collection of data that fully captures the extent of human motor skills.

Human demonstrations also enable data-driven imitation approaches for robot control. It is difficult to formulate a closed-form, rules-based model of how the joint positions of the robot translate to fingertip contact positions and forces on an object. In contrast, data-driven approaches allow for the emergence of dynamic grasping strategies that exploit synergistic redundancies in the robot's kinematics [24], [25]. Providing demonstrations allows a learned controller to bootstrap from human motor skills, rather than relying on random exploration, which can be intractable particularly for high-dimensional systems. Furthermore, a data-driven approach to general-purpose manipulation requires a diverse dataset with good state coverage. Directing demonstrators to achieve multiple tasks across different objects also induces an interesting diversity of grasps and avoids stereotyped behavior in the dataset. However, existing datasets of anthropomorphic robot hands do not contain satisfactory diversity in object affordances and state coverage, or do not feature a realistic robot arm. Similarly, there is a lack of standard benchmarks to compare the performance of anthropomorphic robotic manipulation to human-level performance. A general-purpose anthropomorphic manipulation benchmark would allow the measurement of not only how well artificial agents can imitate specific motor skills, but also the broader capacity for algorithmic generalization.

To tackle these challenges, I turn to the power of human demonstrations to harness

the innate skill of human manipulation through imitation learning. We introduce a novel simulation dataset collected through teleoperation and featuring an anthropomorphic hand and diverse objects. While Chapter 3 will present a method to scale up the algorithmic imitation of demonstrations from diverse domains, this chapter focuses on the considerations for gathering a dataset for imitation.

In this chapter, I present a dataset to enable anthropomorphic imitation, collected via teleoperation of a simulated robot with a direct kinematic mapping to the human hand. It consists of 85 trials of goal-directed manipulation, or approximately 4 hours of data collected at 20 Hz. Each trial consists of goal-directed interaction with objects in a cluttered scene, where the task defining the goal is either bringing an object to a target position and orientation, or object stacking. Observation modalities include robot proprioception, such as joint positions and velocities, multiple high-resolution camera views, ground truth object identities and poses, and fingertip touch sensors. I analyze the intrinsic dimensionality of the dataset by studying the explained variance ratio of a principal component analysis (PCA) over the entire dataset as well as individual subsets for particular objects. This work also investigates the shared dynamics of the robot across the different tasks and objects by training neural networks to predict the future states of the robot and evaluating their prediction error on trajectories including interactions with held-out objects. I benchmark the viability of existing methods for model learning and study how the inclusion of different combinations of inputs and input normalization influences the prediction error for individual observations.

Our analysis implies that a diverse and realistic set of objects induces a complex array of manipulation strategies. This complexity is enabled by the high-dimensional control space of an anthropomorphic hand, and serves as a benchmark challenge for algorithms that model the environment or imitate the demonstrated actions. I believe that datasets like the one presented can motivate future research into imitation learning to solve the problem of dexterous manipulation.

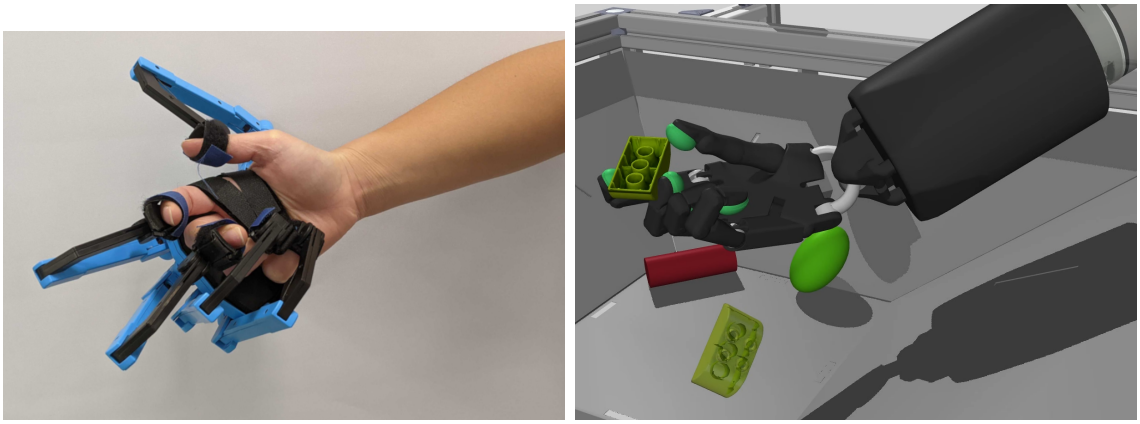

Figure 2.1: Top: A human demonstrator teleoperating the Shadow Hand with the SenseGlove. Bottom: The simulated Shadow Hand manipulating an object to achieve the visualized goal pose.

2.2 Related Work

Classical approaches to multi-finger grasping include force closure criteria for stability [26], [27] and optimization-based finger positioning [28]–[31]. Force closure methods have also been extended to consider task constraints [32]. However, these approaches have various limitations: they make assumptions due to friction models, are not robust to uncertainty, and have no adaptation capability to new objects.

The data-driven study of human motor skills has been shown to aid the design of robot grasping algorithms. Early work focused on reducing the inherent dimensionality of the human hand to design simpler and more efficient planners for high-dimensional robot hands [33]. The postural synergies of the hand were also analyzed from the perspective of optimal grasping forces [34]. The GRASP Taxonomy quantified common human grasps according to the force exerted and the mass and shape of the grasped object [35]. [36] showed that valuable abstractions can be transferred from human grasp demonstrations to robot hands, and [37] similarly derived force-compliant grasping from human demonstrations.

A data-driven approach to extracting abstract motor skills and perceptual representations from human data requires a large-scale dataset. To make human data useful for a specific robot, a mapping from the morphology of the human hand(s) collecting the data to a specific robot hand must be found. Teleoperation of dexterous robotic hands may involve instrumentation

of the hand via a glove or visual markers. Markerless hand tracking techniques allow for unconstrained user movements and may have lower cost to distribute, but is less accurate than instrumented teleoperation. For example, HOannotate is a computer vision method that annotates images of human hands with pose skeletons [38]. However, the geometric transformation from the HO3D demonstrator's hand shape and the coordinate system of the data to a robot is not obvious. One method to bootstrap a markerless hand tracking system from instrumented visual data is DexPilot [39], which resulted in impressive precision on fine motor tasks with a multi-finger robot hand. However, no dataset from this work was publicly released. To collect my dataset, the teleoperators wore an exoskeletal glove with a 6-DOF pose sensor attached to the back of the hand. This method trades off the benefits of markerless teleoperation for much higher precision and dexterity of the resulting motions.

My dataset is not the first robotics dataset to feature a high-dimensional anthropomorphic hand. The CMTouch dataset [40] features the same robot as my dataset interacting with a variety of objects and offers a rich set of observations, including touch and vision, but does not include rewards for specific tasks or a moving arm, limiting the robot's manipulation capabilities. The Adroit domain of D4RL [41] features a floating robot hand accomplishing several object manipulation and tool use tasks. It was shown to be useful for learning from demonstrations with online fine-tuning with DAPG [42] as well as offline reinforcement learning baselines. In contrast, my dataset features a realistic robot arm, which enables the future possibility of transferring learned motor skills to the real hardware. The wider range of sensory modalities and objects also makes it suitable for studying task transfer and object affordances. Finally, my dataset is an order of magnitude larger, which is more conducive to deep learning approaches.

2.3 Dataset

We designed the dataset for the following criteria:

Rich, cross-modal observations. By observing as much sensory information as possible, deep learning models can correlate perception with control and informatively compress the high-dimensional space through representation learning via self-supervision [43].

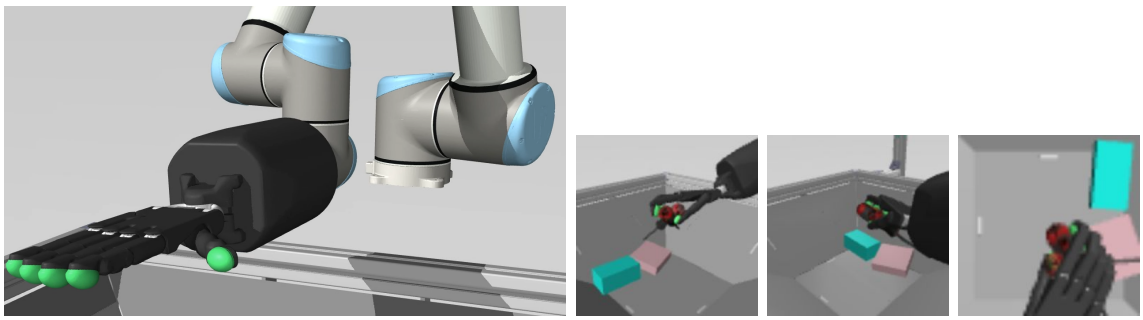

Figure 2.2: Top: The deterministic start state of the robot at each teleoperation trial. Bottom: the available camera views.

Diverse affordances. Real-world manipulation spans a vast variety of object shapes and sizes. A dataset of many objects lends itself to studying the common representations for manipulating diverse objects.

Ground truth rewards. To apply techniques such as offline reinforcement learning to extract goal-oriented behaviors, a reward signal is required, as well as examples that achieve a range of rewards.

Good state coverage. A dataset that spans the extent of possible observations will mitigate overfitting and allow comprehensive description of the environment.

2.3.1 Robot environment

The environment is simulated in MuJoCo [44] and resembles an existing real-world setup. It features a simulated model of the Shadow Dexterous Hand [45] attached to a 6-DOF UR10 arm. The hand has 24 DOFs and 20 degrees of actuation; redundant joints are coupled via tendons. Each of the five fingers is equipped with a touch sensor at the tip with a 3-dimensional reading representing the contact normal force and tangential friction forces, where the magnitude of each vector represents the force quantity in Newtons. The hand and arm are controlled by velocity actuators, thus inputs from the teleoperator must be converted into valid velocity commands in the robot's action space. At the start of each trial, the arm and hand are set to a deterministic state pictured in Fig. 2.2. The arm is mounted at a basket containing various objects. Three camera views at a resolution of 256 px $\times$ 256 px are available: two at the front corners of the basket and one overhead.

The recorded observations are joint position and velocity for all hand and arm joints,

camera views, joint space and Cartesian actions, wrist position and velocity, touch sensors, fingertip and palm positions, object position and orientation, task-specific goal, object identity, and ground truth rewards.

2.3.2 Teleoperation apparatus

The teleoperation system uses a wearable exoskeleton, the SenseGlove [46], to track the 3D position and orientation of the operator's fingertips (see Fig. 2.1). Because the glove tracks fingertip poses relative to the base of the thumb, a proprietary 6-DOF pose sensor is attached on top of the operator's palm. The origin of the pose sensor on the palm is fixed to the workspace, allowing the operator's hand to be tracked whilst moving freely through space. Since there is a kinematic mismatch between the Shadow Hand and the operator's hand, I compute scale and offset factors for each fingertip using least-squares minimization such that the initial sensor positions match the simulated hand's fingertip positions after scaling and offset are applied.

I solve six independent constrained quadratic programs (one for each finger, and one for the palm) to convert the measured Cartesian poses from the teleoperation system to instantaneous joint velocity commands, similar to [47]. An off-the-shelf QP solver is used to solve the resultant optimization problems [48], [49]. Constraints are used to add joint position/velocity limits and collision avoidance between the arm and the basket [50].

2.3.3 Object set

At the start of each trial, three objects are spawned in the basket, with properties described by one of three scenarios:

Primitive objects. One primary object (box, pen, or egg-shaped) is spawned at a deterministic pose in the center of the basket, and two "clutter" objects with random shapes, dimensions and colors are spawned randomly in non-colliding poses in the workspace.

YCB objects. The primary object is sampled from a subset of 17 objects from the YCB Object Set [51], and the clutter is the same as above. Some of the objects have visual rotational symmetries, which makes achieving specific orientations difficult for demonstrators

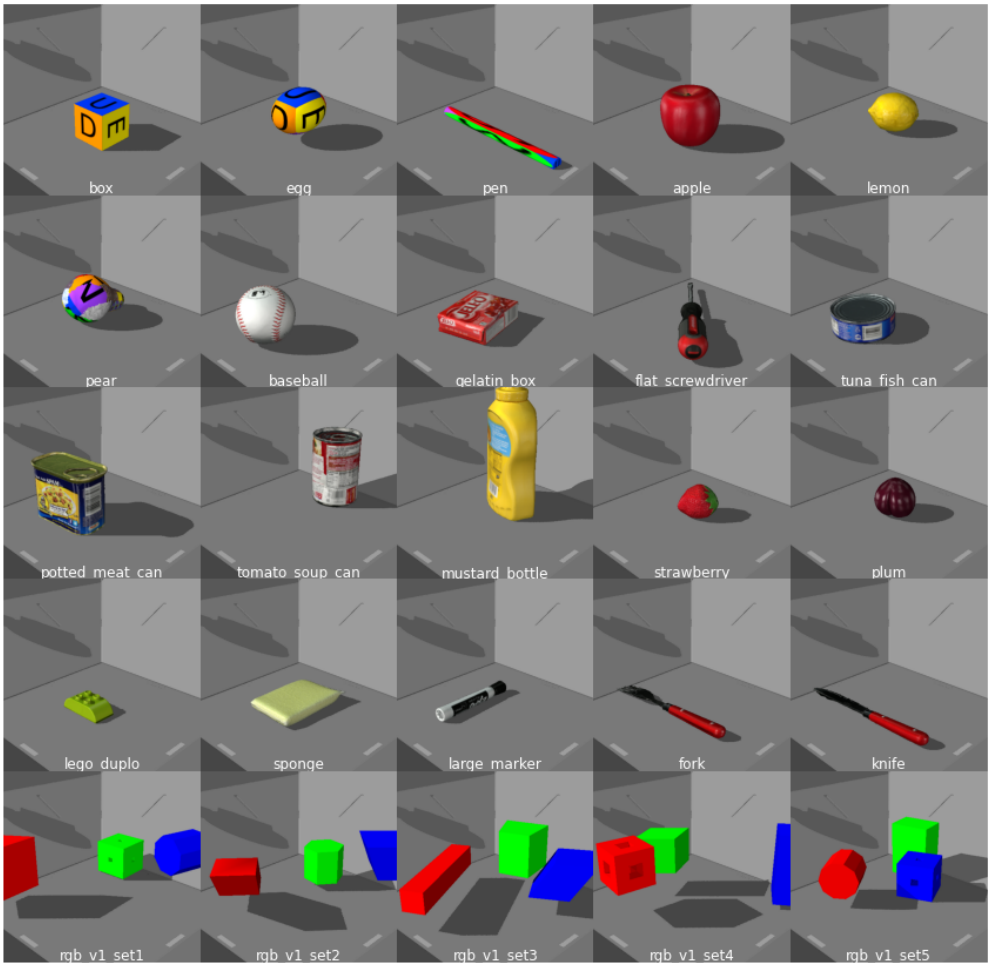

Figure 2.3: All 25 objects manipulated in the dataset. Objects used for clutter are not shown. The pear object is shown with the orientable texture applied to symmetric objects to help the demonstrators disambiguate the object's orientation (second row, first column)).

(strawberry, plum, apple, lemon, pear, sponge, duplo). Thus, I asymmetrically re-texture the symmetric objects as shown in Fig. 2.3.

RGB objects. Three objects from the RGB-Stacking object dataset [52] are spawned at random poses in the workspace.

2.3.4 Rewards

I design goals based on two tasks: object positioning/reorientation and stacking. To quantify how well a demonstrator is achieving a given task, I define reward functions for each task. I compute a combined shaped reward that accumulates as successes are achieved. Specifically,

$$r_t = N\alpha + \Delta r_t, \tag{2.1}$$

where r_t is the accumulated reward, $\Delta r_t \in [0, 1]$, and N is the number of times previous values of Δr_t exceeded success threshold α. Task-specific rewards Δr_t are defined below.

Go To Target Pose: For the primitive and YCB objects, the demonstrator is shown a goal pose, visualized as a transparent "ghost" object (seen in Fig. 2.1). After the goal is achieved, a new goal is sampled randomly: the 3D position is sampled uniformly within the workspace bounding box, and the orientation is sampled from a uniform random distribution. For object position $\vec{p}_t$, object orientation as a quaternion $\vec{q}_t$, goal position and orientation $\vec{p}_t^{\,g}$, $\vec{q}_t^{\,g}$, bounding radius for the goal b, shaping margin parameter σ, we define

$$\phi(\chi) = \begin{cases} 1 & \text{if } |\chi| < b \\ \exp\left(-\frac{(\chi-b)^2}{2\sigma^2}\right) & \text{otherwise} \end{cases}, \quad \chi \in \mathbb{R} \tag{2.2}$$

$$\Delta r_t = \phi(\|\vec{p}_t - \vec{p}_t^{\,g}\|) + \phi(\|\vec{q}_t \vec{q}_t^{\,g-1}\|) \tag{2.3}$$

Stacking: For the RGB objects, the demonstrator is told to stack a certain block on top of another block, such as "stack red on green". Goals are picked from every combination of pairs from the triplet. A sparse reward is given when the object is stacked on the other and then a new target is given. Let $\vec{a}_t$ be the coordinates of the object to stack on top and $\vec{b}_t$ be the coordinates of the desired bottom object. z_{min} is the minimum distance between the

objects when $\vec{a}_t$ is on top, and tol is the maximum distance between the xy coordinates of the objects. Then,

$$\Delta r_t = \begin{cases} 1 & \text{if } |a_{tz} - b_{tz}| > z_{\min} \text{ and } \|\vec{a}_{txy} - \vec{b}_{txy}\| < \text{tol} \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

is the task-specific reward used in Eq. (2.1). This reward is not a precise indicator of success; it assigns success to a top object that is aligned in the xy plane but hovering above the bottom object. Demonstrators were instructed to make complete and stable stacks rather than focus solely on achieving reward.

2.3.5 Data collection procedure

Before data collection, each demonstrator has a 1–3 minutes trial period of undirected play with the box, to become familiar with teleoperation. Then the demonstrator is instructed to manipulate the various objects to achieve a particular goal. Each episode ends after 3–4 minutes, and terminates early if the primary object falls outside of the workspace or the Cartesian controller fails to find a solution. At the end of an episode, the arm resets to its original state, and different objects are spawned into the workspace.

The dataset consists of 83 trajectories from seven demonstrators, totalling to 300,000 timesteps at 20 Hz (the control rate of the robot) or approximately 4 hours and 10 minutes; see Fig. 2.4 and Fig. 2.5 for bar charts illustrating the composition of demonstrators and objects and a comparison of the respective average rewards.

2.4 Analysis

2.4.1 Characterizing grasp pose diversity

To verify if my design decisions succeeded in encouraging diversity, I analyzed the inherent dimensionality of the dataset via principal component analysis (PCA).

I computed a PCA over the hand joint positions across all trajectories, as well as subsets

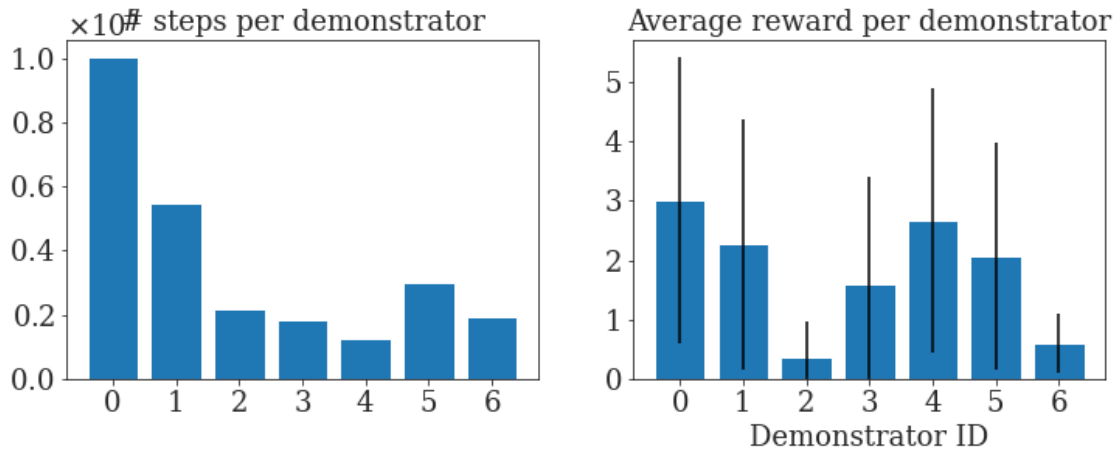

Figure 2.4: Left: Number of timesteps per demonstrator, anonymized by integer index. Right: Reward per demonstrator averaged over all timesteps. Error bars indicate standard deviation. Demonstrators contributed non-uniformly to the dataset, but amount of contribution was not necessarily proportional to performance in terms of measured reward.

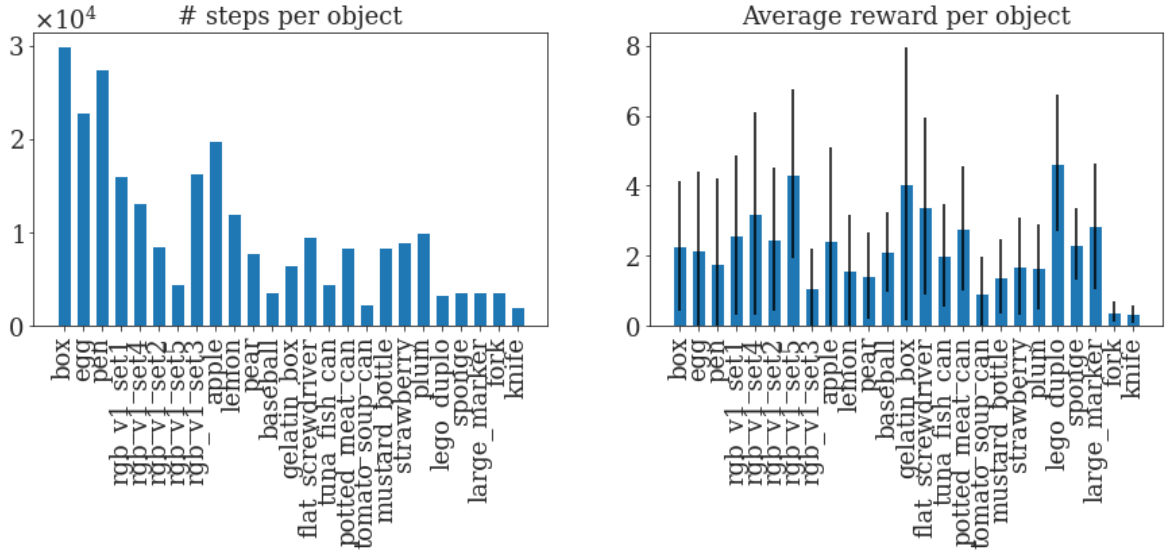

Figure 2.5: Left: Number of timesteps per object. Right: Reward per object averaged over all timesteps. There is a non-uniform representation of objects in the dataset, but frequency does not necessarily correlate to reward (for example, box has the most examples but a middling reward average, and duplo has few examples but one of the highest reward averages).

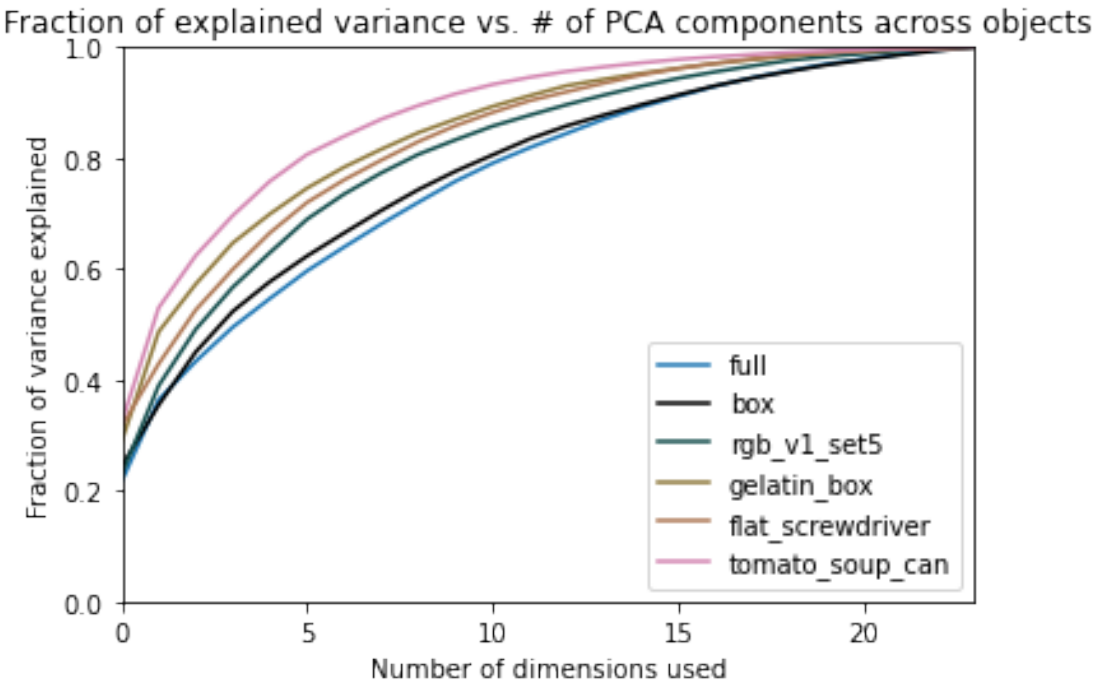

Figure 2.6: The fraction of explained variance as a function of the number of PCA components used. Each curve represents a subset of the data containing only one object. Explaining 95% of the variance requires $\tilde{1}6$ principal components for simplest subset of the data, and 26 (the full dimension) for the full dataset, suggesting that manipulating a wide set of objects required diverse grasping strategies.

containing interactions with only one object, and compared how the cumulative fraction of explained variance varied over components. In Fig. 2.6, curves that asymptotically approach 1 sooner correspond to data subsets that are explainable with fewer components, and therefore contain simpler movements. Curves that are more linear correspond to data subsets that require more principal components to explain the variation.

Because the amount of data varies per object, and the amount of data available influences the diversity of the data, I also considered the relationship between the explained variance ratio and the number of timesteps used for the PCA in Fig. 2.7 for a fixed number of five components. Data points in the lower left part of the scatterplot (such as plum and strawberry) have a high amount of diversity for a relatively small amount of data, while points toward the upper right (such as pen) represent objects that are more common in dataset but induced more stereotyped grasps.

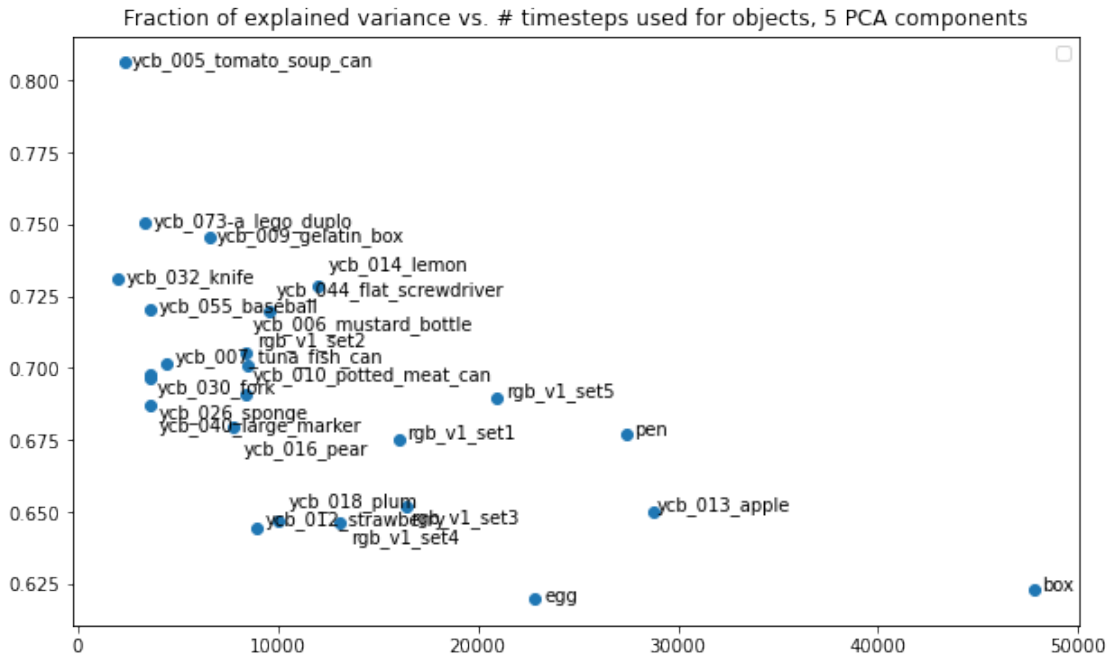

Figure 2.7: For a fixed number of components, the fraction of explained variance for a PCA of a particular object data subset as a function of the size of that subset. If the amount of data per object was the only explanatory factor for the data variance, the plot would be linear. However, the outliers suggest that there are other factors, such as object shape.

2.4.2 Modelling robot dynamics

Model-based reinforcement learning has seen recent success on high-dimensional dexterous manipulation tasks [53]. In particular, model-based offline RL approaches [54], [55] have shown promising results on learning dexterous manipulation from demonstration data [41]. These approaches learn a predictive model of the underlying environment's transition dynamics, predicting future states from the current states and actions. Pre-training a dynamics model from demonstration data can also be beneficial for control and planning because it exposes parts of the state-action space that might be intractable for random exploration to discover, as studies of model-based offline reinforcement learning have shown [56], [57]. To encourage the application of such approaches to my dataset, I benchmark different methods for learning dynamics models on my dataset.

To test if dynamics prediction generalizes across objects, I partitioned the data into a training and test set based on object type. All trajectories that included mustard bottle, sponge, marker and Duplo were partitioned into the test set, resulting in an 85/15% split between training and test data. I split the episodes into fixed-length sequences of length 50, and further sample sub-sequences of length 10 when batching the data. I also zero-centered and scaled the data by computing the mean and standard deviation in each dimension and scaling an example x such that $x' = \frac{x-\mu}{3\sigma}$. I refer this to normalization because the expected range of the data after scaling is $(-1, 1)$.

The input states to the network are joint positions and velocities for the hand and arm, position and velocity of the wrist, fingertip positions, touch sensors, object centroid position and orientation, and the joint space action at that timestep. The network outputs a next-step prediction of each of these states.

All input state dimensions and the action are concatenated together, then layer normalization is applied. My network architecture is a 2-layer MLP with an elu non-linearity, where each layer is of size 128. The outputs of the MLP are then integrated forward together with the previous step. Additive euler integration with $\Delta t = 0.05$ is used for all states except for the orientations, which are integrated using quaternion multiplication, resulting in a quaternion output.

For state $\vec{x}_i$ at timestep i, action $\vec{u}_i$, neural network parameters $\vec{\theta}$ and dataset $\mathcal{D}$ with size $N_{\mathcal{D}}$, I compute the predicted state at time $i+1$ as

$$\vec{\hat{x}}_{i+1} = \vec{x}_i + \Delta t\, f(\vec{x}_i, \vec{u}_i, \vec{\theta}), \tag{2.5}$$

where the neural network f models the change of state between two time steps. I consider three ways of training the neural network:

Single-step loss. I train the neural network to minimize the mean squared error loss

$$\vec{\theta}^* = \arg\min_{\vec{\theta}} \frac{1}{N_{\mathcal{D}}} \sum_{\vec{x},\vec{u}\in\vec{\mathcal{D}}} \|\vec{x}_{i+1} - \vec{\hat{x}}_{i+1}\|^2 \tag{2.6}$$

between the predicted observation and the target observation for each step in the fixed-length subsequence.

Multi-step loss. As an alternative to the single-step loss in Eq. (2.6), we can also train the network by minimizing the multi-step (trajectory) loss

$$\vec{\theta}^* = \arg\min_{\vec{\theta}} \frac{1}{H N_{\mathcal{D}}} \sum_{\vec{\mathcal{D}}} \sum_{i=1}^{H} \|\vec{x}_{t+i} - \vec{\hat{x}}_{t+i}\|^2 \tag{2.7}$$

over an H-step horizon, where $\vec{\hat{x}}_{t+i}$ is a i-step ahead prediction using Eq. (2.5), starting from $\vec{\hat{x}}_t = \vec{x}_t$. The multi-step loss Eq. (2.7) makes sure that long-term predictions using the neural network are consistent with the data, a property that the single-step loss in Eq. (2.6) cannot account for.

Negative log-likelihood loss. I also tested a single-step probabilistic model similar to PETS [58], which estimates the uncertainty of the prediction. This model is trained by minimizing the negative log-likelihood in this estimated distribution.

I trained the model over different configurations with 3 random seeds each, varying single-step vs. multi-step prediction, raw data vs. normalization, and probabilistic vs. deterministic prediction. In all cases, the loss computed over the test set was the single-step loss of Eq. (2.6). For the probabilistic model, the mean prediction was used for evaluation error.

Across all states, the probabilistic model has lower variance across seeds but is not

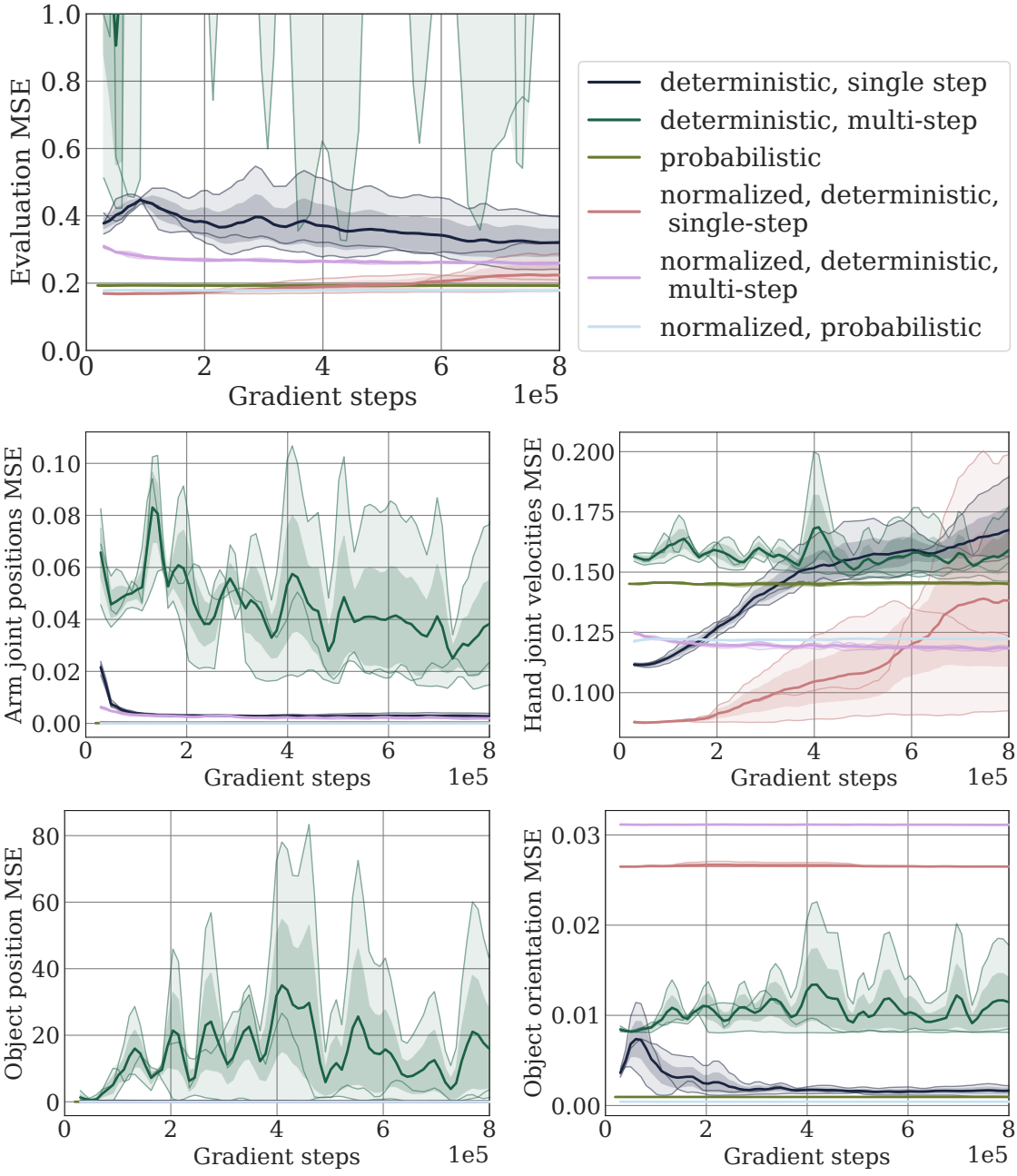

Figure 2.8: Single-step dynamics prediction error computed on the test set for various states as a function of training steps. The un-normalized predictors exhibit extreme overfitting.

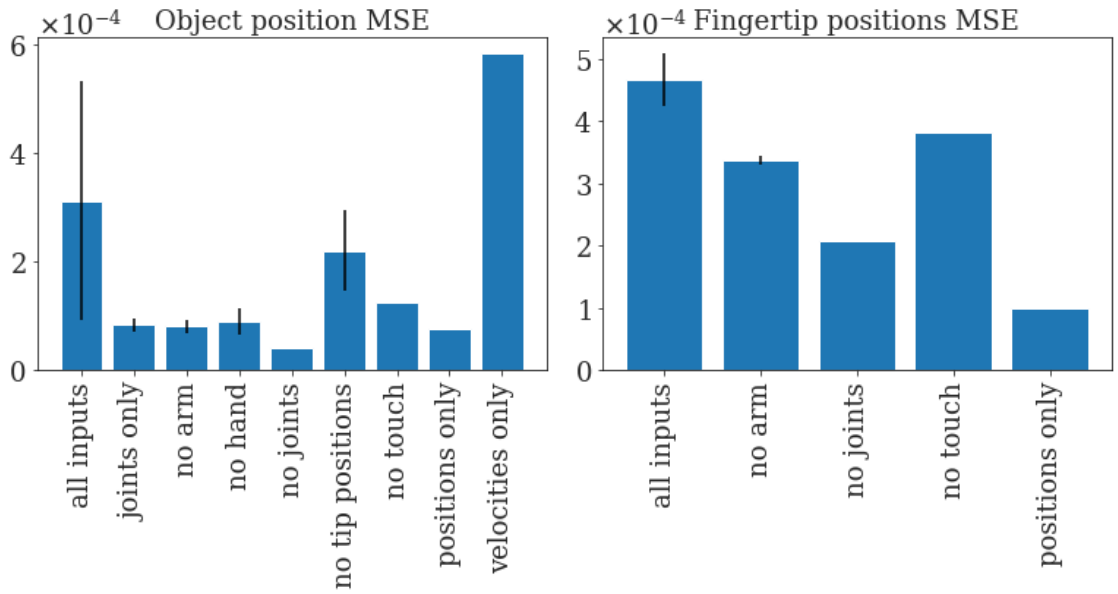

Figure 2.9: Left: Dynamics prediction MSE on the evaluation set for object position over various combinations of inputs, averaged over 15 samples taken after 700,000 gradient steps. Omitting fingertip positions or only using velocities results in higher error. Right: The same plot for fingertip position prediction. Omitting touch sensors has an adverse effect on prediction.

necessarily more accurate in terms of MSE. The un-normalized probabilistic model also has lower variance across seeds than the un-normalized deterministic models, for which normalization appears to help greatly. The object position MSE increases significantly for the un-normalized multi-step model. The normalized deterministic single-step case is fairly stable and accurate for most states but exhibits overfitting and instability for hand joint velocity error. Normalization for object orientation appears to worsen performance in the deterministic case, while the probabilistic models are the most accurate. This could be because normalization does not account for the orientation's quaternion representation.

I also explored which states to use as input to the prediction network by comparing final test set MSE over different sets of input combinations. Including all of the states from the previous experiment tended to result in higher error for individual observations, indicating that the network capacity might not be large enough to predict all states accurately. In Fig. 2.9, I visualize the final evaluation MSE per input combination for object position and fingertip position error. For the object position prediction, omitting fingertip positions results in high error, while omitting arm and hand joint position and velocities results in lower

error, indicating that the network predicts the object state directly from fingertip positions and does not make use of joint information. Omitting all position states and only using velocities is, intuitively, not enough to predict object position accurately. On the other hand, when examining fingertip position error, omitting touch sensor inputs results in high error, suggesting that the contact events at the fingertips detected by the touch sensors are informative of where the fingertips will go next.

These findings show the importance of correct normalization of the data in model learning. They also suggest that probabilistic models exhibit less variance and overfitting on unnormalized data, but make worse predictions on average.

2.5 Conclusion

I introduced a dataset containing 4 hours of trajectories featuring goal-directed object manipulation by the Shadow Hand and UR10 robot arm, collected via teleoperation. I analyzed the complexity and intrinsic dimensionality of the dataset through PCA, showing that a diverse set of grasping strategies was necessary to manipulate the range of shapes present in the dataset. I also examined an important application of data by training dynamics prediction networks and evaluating their prediction error on a held-out set of objects. The results from these experiments confirm existing findings from model learning literature and emphasize the interaction between observations and importance of proper normalization in this dataset.

Model learning – that is, training a network to predict future states in the environment – is a popular approach in robotics when combined with model-based control or model-based reinforcement learning. While imitation learning and model learning are distinct methods, model learning can enable forms of imitation learning through model-based planning to achieve goal states. Additionally, the ability to predict future states lays the foundation for more complex cognition and behavior.

Learning to predict future states in the environment could be instrumental to the overarching goal of human imitation even without the direct use of imitation learning, as it is a foundational ability for more complex cognition [59]. While this chapter focused on model

learning for a single robotic agent to manipulate objects in an environment, model learning could also apply to predicting the actions of other agents. This theme is discussed further in Chapter 4, in which the goal is to imitate a more human-like representation of social identity, which includes an aspect of modelling the minds and behaviors of others.

This chapter also investigated how certain kinds of errors occur in the dynamics model. Model error can be an indicator, or perhaps a catalyst, of emergence. When reality differs from the expectation of a machine learning system, the source of discrepancy must be investigated – and sometimes revealed to have unfolded from an immeasurable complexity. The theme of emergence and how imitation of certain qualities can emerge even when it is not the explicit goal of a system is further explored in Chapter 5.

Notably, this chapter did not explicitly demonstrate imitation learning as a method. That is, I did not use the demonstration data to train a policy for outputting actions for an agent to mimic the actions recorded by teleoperators. Due to the multi-object, multi-task nature of the data, the method for training an imitation policy must be capable of transfer learning and robust to distribution shift. The next chapter presents a suitable technique for learning imitation from diverse demonstration data that spans not only multiple tasks, but also multiple embodiments.

Chapter 3

A Generalist Agent

Given a diverse dataset of demonstrations, a robust and scalable method for imitation learning is needed to reproduce these recorded behaviors in a neural network. Inspired by progress in large-scale language modeling, we applied a similar approach towards building a single generalist agent beyond the realm of text outputs. The agent, which we refer to as Gato, works as a multi-modal, multi-task, multi-embodiment generalist policy, trained via an imitation learning rule. The same network with the same weights can play Atari, caption images, chat, stack blocks with a real robot arm and much more, deciding based on its context whether to output text, joint torques, button presses, or other tokens. In this chapter, we emphasize Gato's performance on out-of-distribution, real-world robotics tasks. We also present visualizations of Gato's internal activations and embedding space to aid the interpretation of its capacity to generalize.

3.1 Introduction

This chapter completes our investigation of imitation as a method by presenting an imitation learning system. The dataset described in Chapter 2 is a first step towards engineering a general-purpose anthropomorphic manipulation system, capable of manipulating diverse objects and accomplishing multiple tasks. The next step is to develop a machine learning method which can effectively train an imitation policy from this data. One of the aims of gathering a dataset for general-purpose manipulation is to imitate the human capacity for

generalization–the ability to recognize patterns and meaningfully apply acquired skills in new circumstances. We are not only interested in training a model which can reproduce specific behaviors in different modalities from a highly diverse dataset (generality), but also exhibits this capacity for rapid adaptation (generalization). We call a system which can exhibits both generality and generalization while acting across a wide range of environments, including novel ones, a *generalist agent*

While the dataset of the previous chapter encapsulates a diverse set of goals and behaviors, many imitation techniques in robotics are highly specialized and optimized for a singular objective. There are many compounding difficulties associated with training a general-purpose robotic agent. Gathering high-quality demonstration data for robotics is expensive and time-consuming. Classical architectures such as multi-layer perceptrons and recurrent neural networks have known limitations when it comes to generalization and catastrophic forgetting. Furthermore, representing the many possibilities of multi-sensory world (and simulations thereof) presents the challenges of representation and encoding: how to tabulate or serialize the rich, diverse, and high-dimensional environmental data without aliasing away key details.

However, an agent that learns from a massive volume of diverse data could potentially synthesize common representations across tasks and domains and then leverage these findings to generalize to new settings. This representational transference could ease the difficulties around gathering robotics datasets, by allowing the agent to leverage data from domains outside of robotics. To that end, this chapter investigates building a generalist agent that can imitate experts with different embodiments, sensory views, and affordances. In order to address the limitations described above, the model architecture should facilitate generalization, and the data must follow a domain-agnostic serialization scheme. Although this agent has not yet been trained on the dataset of the previous chapter, it could be extended to do so. While Chapter 2 focused on data recorded from a single robotic embodiment, presenting a large-scale model with training data derived from many different kinds of agents could enable cross-embodiment transfer of skills and behaviors.

The aim of this chapter is to build and evaluate a large transformer sequence model trained to imitate expert behavior on a large variety of tasks. The architecture builds off of prior work in large-scale language modelling contemporary to this work, such as GPT-3

[60] and Gopher [61]. Unlike predecessor agents, which are specialized to a closed set of environments, Gato operates on diverse data: its inputs and outputs can be anything that can be serialized into a flat sequence. This enables the agent to not only act in a variety of environments, but to also control different embodiments. The hypothesis behind this architecture is that performance will improve with further scale [62], [63], and broadening the domain of environments increases the availability of potential training data. Furthermore, the system can adapt to new environments and tasks with a small amount of data. The transformer architecture's inductive bias for natural language and the presence of natural language in the training data can serve as a symbolic grounding across tasks [64], aiding with the challenge of transfer.

As originally shown in Reed, Zolna, Parisotto, *et al.* [1], Gato was a proof of concept for the hypothesis that scaling up imitation on multi-embodiment, multi-task demonstrations is "all you need". While I discuss some of the technical limitations of this statement in Section 3.7, there are also greater philosophical and ethical issues associated with such sweeping statements [65]. However, Gato demonstrated that, with a flexible encoding scheme and scalable transformer architecture, it is possible to deploy foundation models on real robotic hardware and use their few-shot learning capabilities for out-of-distribution transfer learning.

3.2 Model

Gato is trained on a wide variety of data featuring diverse modalities such as images, text, robotics sensor data, and discrete and continuous action inputs. All data is processed into a flat sequence of discrete tokens, which enables the model to process inputs of various shapes and sizes from different training environments.

To enable processing this multi-modal data, data is serialized into a flat sequence of tokens. In this representation, Gato can be trained and sampled from akin to a standard large-scale language model. During deployment, sampled tokens are assembled into dialogue responses, captions, button presses, or other actions based on the context. In the following subsections, I describe Gato's tokenization, network architecture, loss function, and deployment.

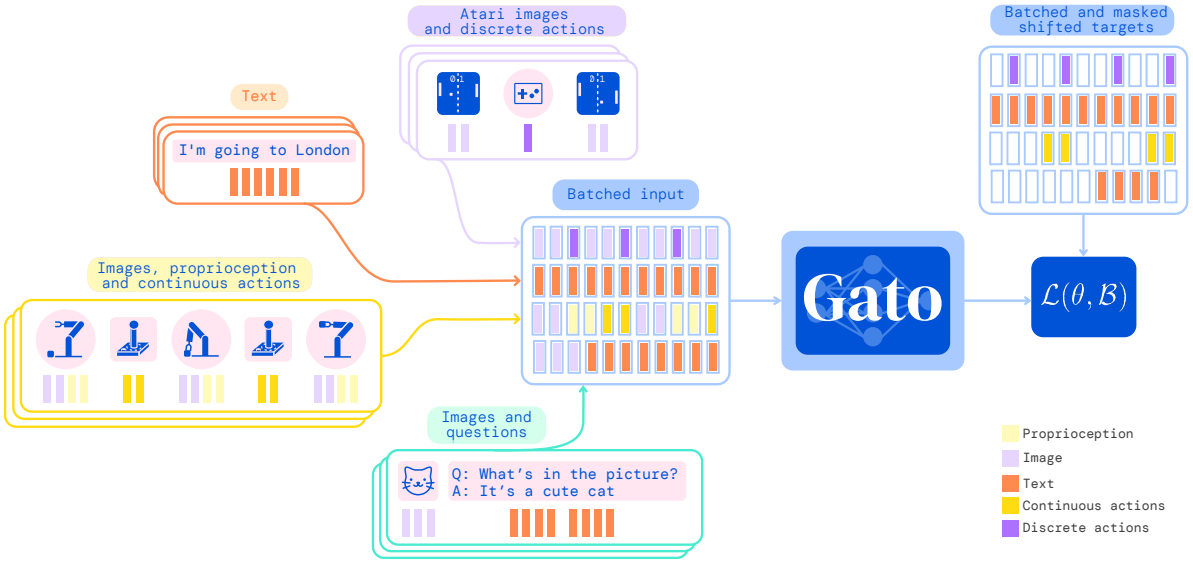

Figure 3.1: Training phase of Gato. Data from different tasks and modalities is serialized into a flat sequence of tokens, batched, and processed by a transformer neural network akin to a large language model. Masking is used such that the loss function is applied only to target outputs, i.e. text and various actions.

3.2.1 Tokenization

There are infinite possible ways to transform data into tokens, including directly using the raw underlying byte stream. Below is the tokenization scheme which produced the best results for Gato at the current scale using contemporary hardware and model architectures.

- Text is encoded via SentencePiece [66] with 32000 subwords into the integer range $[0, 32000)$.
- Images are first transformed into sequences of non-overlapping 16×16 patches in raster order, as done in ViT [67]. Each pixel in the image patches is then normalized between $[-1, 1]$ and divided by the square-root of the patch size (i.e. $\sqrt{16} = 4$).
- Discrete values, e.g. Atari button presses, are flattened into sequences of integers in row-major order. The tokenized result is a sequence of integers within the range of $[0, 1024)$.
- Continuous values, e.g. proprioceptive inputs or joint torques, are first flattened into

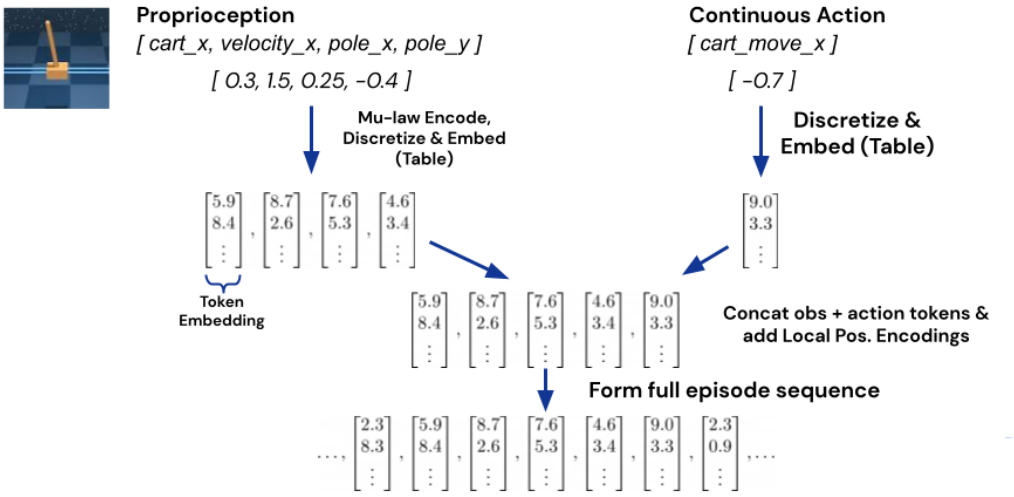

Figure 3.2: A visualization of tokenizing and sequencing continuous values, e.g. proprioception.

sequences of floating point values in row-major order. The values are mu-law encoded to the range $[-1, 1]$ if not already there (see Figure 3.2 for details), then discretized to 1024 uniform bins. The discrete integers are then shifted to the range of $[32000, 33024)$.

After converting data into tokens, the following canonical sequence ordering is used.

- Text tokens in the same order as the raw input text.
- Image patch tokens in raster order.
- Tensors in row-major order.
- Nested structures in lexicographical order by key.
- Agent timesteps as observation tokens followed by a separator, then action tokens.
- Agent episodes as timesteps in time order.

3.2.2 Embedding input tokens and setting output targets

After tokenization and sequencing, a parameterized embedding function $f(\cdot; \theta_e)$ is applied to each token (i.e. it is applied to both observations and actions) to produce the final model

input. To enable efficient learning from the multi-modal input sequence $s_{1:L}$ the embedding function performs different operations depending on the modality the token stems from:

- Tokens belonging to text, discrete- or continuous-valued observations or actions for any time-step are embedded via a lookup table into a learned vector embedding space. Learnable position encodings are added for all tokens based on their local token position within their corresponding time-step.
- Tokens belonging to image patches for any time-step are embedded using a single ResNet [68] block to obtain a vector per patch. For image patch token embeddings, a learnable within-image position encoding vector is added.

As the agent models data autoregressively, each token is potentially also a target label given the previous tokens. Text tokens, discrete and continuous values, and actions can be directly set as targets after tokenization. Image tokens and agent nontextual observations are not currently predicted in Gato, although that may be an interesting direction for future work. Targets for these non-predicted tokens are set to an unused value and their contribution to the loss is masked out.

3.2.3 Training

Gato is trained on a total of 63 million episodes from 596 separate tasks, plus the vision-language datasets used in [69]. The tasks include simulated DeepMind Control Suite tasks in MuJoCo, DeepMind Lab tasks, Atari games, simulated robotics domains include Metaworld, DeepMind Manipulation Suite, and the RGB Stacking benchmark training dataset, as well as real robot trajectories from the RGB Stacking dataset.

Given a sequence of tokens $s_{1:L}$ and parameters θ, model the data using the chain rule of probability:

$$\log p_\theta(s_1, \ldots, s_L) = \sum_{l=1}^{L} \log p_\theta(s_l | s_1, \ldots, s_{l-1}), \tag{3.1}$$

Let b index a training batch of sequences $\mathcal{B}$. We define a masking function m such that $m(b, l) = 1$ if the token at index l is either from text or from the logged action of an agent,

and 0 otherwise. The training loss for a batch $\mathcal{B}$ can then be written as

$$\mathcal{L}(\theta, \mathcal{B}) = -\sum_{b=1}^{|\mathcal{B}|} \sum_{l=1}^{L} m\left(b, l\right) \log p_\theta \left(s_l^{(b)} | s_1^{(b)}, \ldots, s_{l-1}^{(b)}\right) \tag{3.2}$$

Thus the goal of optimization is to predict the next likely action given the previous tokens, which is a sequence modelling problem. Given their impressive sequence modelling capabilities, we chose a 1.2B parameter transformer to predict the distribution over the next token in a sequence of environment steps.

As described above, Gato's network architecture has two main components: the parameterized embedding function which transforms tokens to token embeddings, and the sequence model which outputs a distribution over the next discrete token. While any general sequence model can work for next token prediction, we chose a transformer [70] for simplicity and scalability. Gato uses a 1.2B parameter decoder-only transformer with 24 layers, an embedding size of 2048, and a post-attention feedforward hidden size of 8196.

To differentiate between different task specifications, prompt conditioning is used: the predicted sequence is prepended to to a randomly sampled subsection from an episode in the task the agent is currently acting. Because distinct tasks within a domain can share identical embodiments, observation formats and action specifications, the model sometimes needs further context to disambiguate tasks. Rather than providing e.g. one-hot task identifiers, prompt conditioning inspired by [60], [71], [72] is used. During training, for 25% of the sequences in each batch, a prompt sequence is prepended, coming from an episode generated by the same source agent on the same task. Half of the prompt sequences are from the end of the episode, acting as a form of goal conditioning for many domains; and the other half are uniformly sampled from the episode. During evaluation, the agent can be prompted using a successful demonstration of the desired task, which we do by default in all results that involve controlling an agent or robotic embodiment.

Training of the model is performed on a 16x16 TPU v3 slice for 1M steps with batch size 512 and token sequence length $L = 1024$, which takes about 4 days. Because agent episodes and documents can easily contain many more tokens than fit into context, we randomly sample subsequences of L tokens from the available episodes. Each batch mixes subsequences

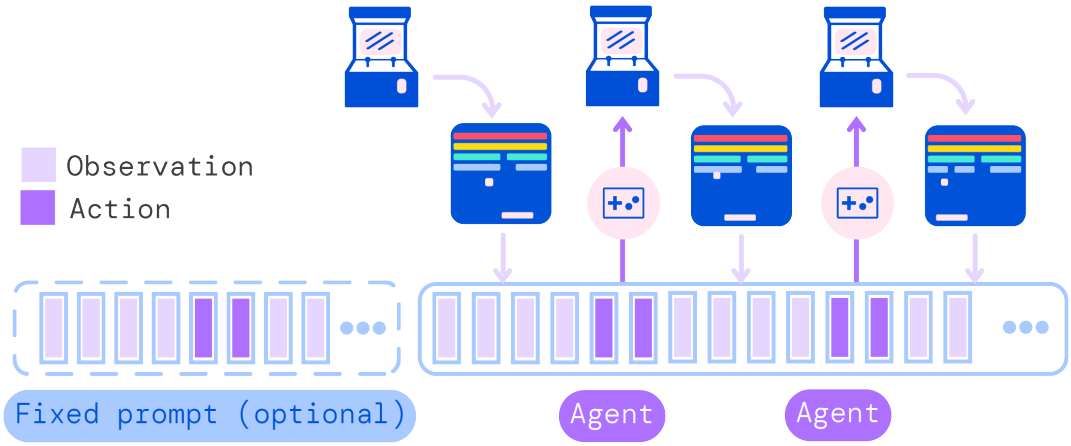

Figure 3.3: Running Gato as a control policy. Gato consumes a sequence of interleaved tokenized observations, separator tokens, and previously sampled actions to produce the next action in standard autoregressive manner. The new action is applied to the environment – a game console in this illustration, a new set of observations is obtained, and the process repeats.

approximately uniformly over domains (e.g. Atari, MassiveWeb, etc.), with some manual upweighting of larger and higher quality datasets (see Table 3.1 in Section 3.3 for details).

3.2.4 Deployment

Deploying Gato as a policy is illustrated in Figure 3.3. First a prompt, such as a demonstration, is tokenized, forming the initial sequence. By default, the first 1024 tokens of the demonstration are taken. Next the environment yields the first observation which is tokenized and appended to the sequence. Gato samples the action vector autoregressively one token at a time. Once all tokens comprising the action vector have been sampled (determined by the action specification of the environment), the action is decoded by inverting the tokenization procedure described in Section 3.2.1. This action is sent to the environment which steps and yields a new observation. The procedure repeats. The model always sees all previous observations and actions in its context window of 1024 tokens. We found it beneficial to use transformer XL memory during deployment, although it was not used during training [73].

Table 3.1: Datasets. Left: Control datasets used to train Gato. Right: Vision & language datasets. Sample weight means the proportion of each dataset, on average, in the training sequence batches.

Control environment	Tasks	Episodes	Approx. Tokens	Sample Weight
DM Lab	254	16.4M	194B	9.35%
ALE Atari	51	63.4K	1.26B	9.5%
ALE Atari Extended	28	28.4K	565M	10.0%
Sokoban	1	27.2K	298M	1.33%
BabyAI	46	4.61M	22.8B	9.06%
DM Control Suite	30	395K	22.5B	4.62%
DM Control Suite Pixels	28	485K	35.5B	7.07%
DM Control Suite Small	26	10.6M	313B	3.04%
DM Control Suite Large	26	26.1M	791B	3.04%
Meta-World	45	94.6K	3.39B	8.96%
ProcGen	16	1.6M	4.46B	5.34%
RGB Stacking simulator	1	387K	24.4B	1.33%
RGB Stacking real robot	1	15.7K	980M	1.33%
Modular RL	38	843K	69.6B	8.23%
DM Manipulation Playground	4	286K	6.58B	1.68%
Playroom	1	829K	118B	1.33%
Total	596	63M	1.5T	85.3%

Vision / language dataset	Sample Weight
MassiveText	6.7%
M3W	4%
ALIGN	0.67%
MS-COCO Captions	0.67%
Conceptual Captions	0.67%
LTIP	0.67%
OKVQA	0.67%
VQAV2	0.67%
Total	14.7%

3.3 Datasets

Gato is trained on a large number of datasets comprising agent experience in both simulated and real world environments, as well as a variety of natural language and image datasets. The datasets used and their attributes are listed in Table 3.1. The approximate number of tokens per control dataset is computed assuming the tokenization mechanism described in Section 3.2.1.

3.3.1 Simulated control tasks

The control tasks consist of datasets generated by specialist SoTA or near-SoTA reinforcement learning agents trained on a variety of different environments. For each environment we record a subset of the experience the agent generates (states, actions, and rewards) while it is training.

The simulated environments include Meta-World [74] introduced to benchmark meta-reinforcement learning and multi-task learning, Sokoban [75] proposed as a planning problem, BabyAI [76] for language instruction following in grid-worlds, the DM Control Lab [77] for continuous control, as well as DM Lab [78], designed to teach agents navigation and 3D vision from raw pixels with an egocentric viewpoint. We also include four tasks using a simulated Kinova Jaco arm, as introduced in [79]. The video game environments include the Arcade Learning Environment (ALE) [80] with classic Atari games, the ProcGen environment [81] and Modular RL [82].

It was effective to train on a filtered set of episodes with returns at least 80% of the expert return for the task. The expert return measures the maximum sustained performance that the expert agent can achieve. It is defined as the maximum over the set of all windowed average returns calculated over all the collected episodes for a task:

$$\max_{j \in [0,1,\ldots,N-W]} \left(\sum_{i=j}^{j+L-1} \frac{R_i}{W} \right)$$

where N it the total number of collected episodes for the task, W is the window size, and R_i is the total return for episode i. To obtain accurate estimates, in practice, the window was set W to be 10% of the total data amount or a minimum of 1000 episodes (i.e. $W = \min(1000, 0.1 \times N)$).

3.3.2 Robotics - RGB Stacking Benchmark (real and sim)

As a testbed for taking physical actions in the real world, we chose the robotic block stacking environment introduced by Lee, Devin, Zhou, *et al.* [52]. The environment consists of a Sawyer robot arm with 3-DoF cartesian velocity control, an additional DoF for velocity, and

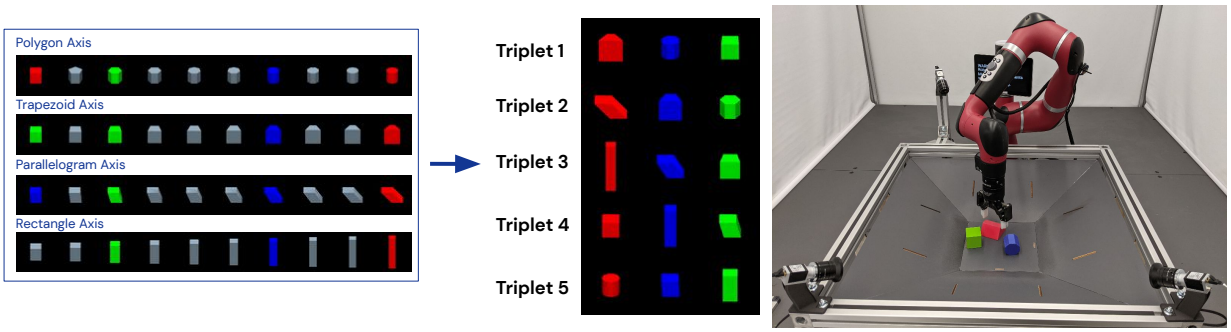

Figure 3.4: RGB Stacking environment with the Sawyer robot arm. Blocks vary along several shape axes, with 5 held out test triplets. The goal is to stack red on blue, ignoring green.

a discrete gripper action. The robot's workspace contains three plastic blocks colored red, green and blue with varying shapes. The available observations include two 128 × 128 camera images, robot arm and gripper joint angles as well as the robot's end-effector pose. Notably, ground truth state information for the three objects in the basket is not observed by the agent. Episodes have a fixed length of 400 timesteps at 20 Hz for a total of 20 seconds, and at the end of an episode block positions are randomly re-positioned within the workspace. The robot in action is shown in Figure 3.4. There are two challenges in this benchmark: *Skill Mastery* (where the agent is provided data from the 5 test object triplets it is later tested on) and *Skill Generalization* (where data can only be obtained from a set of training objects that excludes the 5 test sets).

We used several sources of training data for these tasks. In Skill Generalization, for both simulation and real, we use data collected by the best generalist sim2real agent from Lee, Devin, Zhou, *et al.* [52]. We collected data only when interacting with the designated RGB-stacking *training objects* (this amounts to a total of 387k successful trajectories in simulation and 15k trajectories in real). For Skill Mastery we used data from the best per group experts from Lee, Devin, Zhou, *et al.* [52] in simulation and from the best sim2real policy on the real robot (amounting to 219k trajectories in total). Note that this data is only included for specific Skill Mastery experiments in Section 3.5.4.

In these robotics tasks, we use the sparse reward function described in [52] for data filtering. We only select trajectories with *final* task success; that is, a sparse reward of 1 on the final timestep.

3.4 Related Work

Gato was inspired by works such as GPT-3 [60] and Gopher [61], pushing the limits of generalist language models; and more recently the Flamingo [69] generalist visual language model. Chowdhery, Narang, Devlin, *et al.* [83] developed the 540B parameter Pathways Language Model (PalM) explicitly as a generalist few-shot learner for hundreds of text tasks.

The most closely related architectures to that of Gato are Decision Transformers [84]–[87] and Trajectory Transformer [88], which showed the effectiveness of language model-like architectures for a variety of control problems. Gato also uses an LM-like architecture for control, but with design differences chosen to support multi-modality, multi-embodiment, large scale and general purpose deployment. Pix2Seq [89] also uses an LM-based architecture for object detection. Perceiver IO [90] uses a transformer-derived architecture specialized for very long sequences, to model any modality as a sequence of bytes.

Gato also takes inspiration from recent works on multi-embodiment continuous control. Huang, Mordatch, and Pathak [82] used message passing graph networks to build a single locomotor controller for many simulated 2D walker variants. Kurin, Igl, Rocktäschel, *et al.* [91] showed that transformers can outperform graph based approaches for incompatible (i.e. varying embodiment) control, despite not encoding any morphological inductive biases. Devin, Gupta, Darrell, *et al.* [92] learn a modular policy for multi-task and multi-robot transfer in simulated 2D manipulation environments. Chen, Murali, and Gupta [93] train a universal policy conditioned on a vector representation of robot hardware, showing successful transfer both to simulated held out robot arms, and to a real world sawyer robot arm.

A variety of earlier generalist models have been developed that, like Gato, operate across highly distinct domains and modalities. NPI [94] trained a single LSTM [95] to execute diverse programs such as sorting an array and adding two numbers, such that the network is able to generalize to larger problem instances than those seen during training. Kaiser, Gomez, Shazeer, *et al.* [96] developed the MultiModel that trains jointly on 8 distinct speech, image and text processing tasks including classification, image captioning and translation. Modality-specific encoders were used to process text, images, audio and categorical data, while the rest of the network parameters are shared across tasks. Schmidhuber [97] proposed "*one big net*

for everything", describing a method for the incremental training of an increasingly general problem solver. Keskar, McCann, Varshney, *et al.* [98] proposed controllable multi-task language models that can be directed according to language domain, subdomain, entities, relationships between entities, dates, and task-specific behavior.

To contextualize Gato, we distinguish between one single multi-task network architecture versus one single neural network with the same weights for all tasks. Several poplar RL agents achieve good multi-task RL results within single domains such as Atari57 and DMLab [99]–[101]. However, it is much more common to use the same policy architecture and hyper-parameters across tasks, but the policy parameters are different in each task [102], [103]. This is also true of state-of-the-art RL methods applied to board games [104]. Moreover, this choice has been adopted by off-line RL benchmarks [41], [105] and recent works on large sequence neural networks for control, including decision transformers [84]–[86] and the Trajectory Transformer of [88]. In contrast, in this work we learn a single network with the same weights across a diverse set of tasks.

Our work is based on deep autoregressive models, which have a long history and can be found in generative models of text, images, video and audio. Combining autoregressive generation with transformers [70], [106] has been of enormous impact in language modelling [60], [61], protein folding [107], vision-language models [17], [69], [108], code generation [109], [110], dialogue systems with retrieval capabilities [111], [112], speech recognition [113], neural machine translation [114] and more [115]. Recently researchers have explored task decomposition and grounding with language models [116], [117].

While Gato takes a *tabula rasa* approach by training a single neural network from scratch, other works adapt the weights of a pretrained model into their learning system to provide "off-the-shelf" verbal or sensory capabilities. Li, Puig, Paxton, *et al.* [118] construct a control architecture, consisting of a sequence tokenizer, a pretrained language model and a task-specific feed-forward network. They apply it to VirtualHome and BabyAI tasks, and find that the inclusion of the pretrained language model improves generalisation to novel tasks. Similarly, Parisi, Rajeswaran, Purushwalkam, *et al.* [119] demonstrate that vision models pretrained with self-supervised learning, especially crop segmentations and momentum contrast [120], can be effectively incorporated into control policies.

There has been great recent interest in data-driven robotics [121], [122]. However, Bommasani, Hudson, Adeli, *et al.* [115] note that in robotics, data of sufficient abundance and diversity is the biggest stumbling block to training a foundation model for robotics. Moreover, every time we update the hardware in a robotics lab, we need to collect new data and retrain. This motivates the adaptability of a generalist agent that can adapt to new embodiments and learn new tasks with few-shot data.

3.5 Results

First person teleoperation enables the collection of expert demonstrations. However, such demonstrations are slow and costly to collect. Data-efficient behavior cloning methods are therefore desirable for training a generalist robot manipulator and offline pretraining is thus a well-motivated area of research. To that end, I evaluated Gato on the established RGB Stacking benchmark for robotics.

3.5.1 Real-time considerations for robotics evaluation

In the real world, control is asynchronous; physics does not wait for computations to finish. Thus, inference latency is a concern for evaluating a large model for real world tasks. In robotics, a fast control rate is thought to be critical for reacting to dynamic phenomena. The robot setup for RGB stacking has a 20Hz control rate (0.05 second timestep) by design. In order to reach an acceptable margin of latency, I modified inference at evaluation time by shortening the context length to 1. A parallel sampling scheme was implemented where all the action tokens are zeroed out in the input sequences during training so we can sample all tokens corresponding to a robot action in a single model inference step instead of autoregressively as it's done in other domains.The `1.18B` parameter model was able to run on the hardware accelerators in the robots used (NVidia GeForce RTX 3090s), but still overran the 20Hz control rate by a small amount ($\sim$0.01 seconds).

Table 3.2: Gato real robot Skill Generalization results. In addition to performing hundreds of other tasks, Gato also stacks competitively with the comparable published baseline.

Agent	Group 1	Group 2	Group 3	Group 4	Group 5	Average
Gato	**24.5**%	33%	**50.5**%	76.5%	**66.5**%	**50.2**%
BC-IMP [52]	23%	**39.3**%	39.3%	**77.5**%	66%	49%

3.5.2 Robotics: Skill Generalization

The Skill Generalization challenge from the RGB Stacking robotics benchmark tests the agent's ability to stack objects of previously unseen shapes. The agent is trained on a dataset consisting of episodes of the robot stacking objects with a variety of different shapes. Five triplets of object shapes are, however, not included in the training data and serve as test triplets. I evaluated the trained generalist for 200 episodes per test triplet on the real robot. Table 3.2 shows that the generalist agent's success rate on each test triplet is comparable to the single task BC-IMP (filtered BC) baseline in [52].

3.5.3 Fine-tuning on Robotic Stacking Tasks

Section 3.5.2 demonstrates that the base Gato capable of a diverse array of tasks can perform competitively on the RGB Stacking Skill Generalization benchmark. In this section, the following question is explored: *How does the agent improve on robotics tasks when allowed to fine-tune on new tasks?* I consider different model sizes and analyse the impact of pretraining datasets on the Skill Generalization benchmark, as well as a novel out of distribution task. Further analysis of fine-tuning with dataset ablations is in Appendix 3.5.5.

Skill Generalization

First, I would like to show that fine-tuning on object-specific data, similarly to what was done by Lee, Devin, Springenberg, *et al.* [123], is beneficial. Therefore, I fine-tuned Gato separately on five subsets of demonstrations from the *test* dataset. Each subset was obtained by random partitioning of a test dataset consisting of demonstrations gathered by a generalist sim-to-real agent stacking real test objects. I consider this setting, which is comparable to the fine-tuning baselines on RGB stacking tasks from Lee, Devin, Springenberg, *et al.* [123],

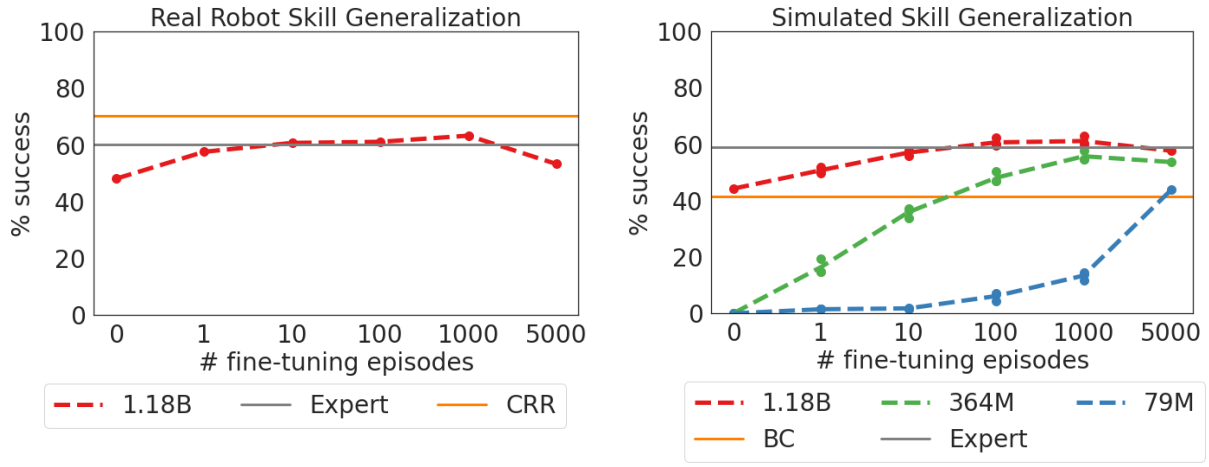

Figure 3.5: Robotics fine-tuning results. Left: Comparison of real robot Skill Generalization success rate averaged across test triplets for Gato, expert, and CRR trained on 35k expert episodes (upper bound). Right: Comparison of simulated robot Skill Generalization success rate averaged across test triplets for a series of ablations on the number of parameters, including scores for expert and a BC baseline trained on 5k episodes.

and use the 5k dataset that their behavior cloning 5k results are obtained with. To best match their experiments, I change the return filtering scheme during training: instead of using only successful stacks, I condition on the normalized return of the episode.

Figure 3.5 compares the success rate of Gato across different fine-tuning data regimes to the sim-to-real expert and a Critic-Regularized Regression (CRR) [124] agent trained on 35k episodes of all test triplets. Gato, in both reality and simulation (red curves on the left and right figure, respectively), recovers the expert's performance with only 10 episodes, and peaks at 100 or 1000 episodes of fine-tuning data, where it exceeds the expert. After this point (at 5000), performance degrades slightly but does not drop far below the expert's performance.

Fine-tuning and Model Size

To better understand the benefit of large models for few-shot adaptation in robotics domains, I conducted an ablation on model parameter size. This section focuses on in-simulation evaluation. Figure 3.5 compares the full `1.18B` parameter Gato with the smaller `364M` and `79M` parameter variants for varying amounts of fine-tuning data. Although the `364M` model overfits on one episode, causing performance to drop, there is a clear trend towards better

adaptation with fewer episodes as the number of parameters is scaled up. The `79M` model performs clearly worse than its bigger counterparts. The results suggest that the model's greater capacity allows the model to use representations learned from the diverse training data at test time.

Adaptation to Perceptual Variations

While the Skill Generalization task is an effective benchmark for motor Skill Generalization to shape variations, it does not test the agent's ability to adapt to perceptual variations and permutations in the objective specification. To further evaluate Gato's generalization capabilities, I devised a new task in the RGB stacking benchmark where the goal is to stack the blue object on the green object, for test triplet 1 (see Figure 3.4). First, I used a 3D mouse to collect 500 demonstrations of this task on the real robot, for a total of 2 hours and 45 minutes of demonstration data, and fine-tuned Gato on these episodes. Notably, all of the simulated and real robotics data in the pretraining set shows the robot successfully stacking the red object on the blue object, and the data does not include the object shapes in the test set. I found that additionally adding simulated demonstrations of the stack blue on green task to the fine-tuning dataset improved performance, and 10% was an ideal sampling ratio for this data.

I achieved a final 60% success rate after evaluating fine-tuned Gato on the real robot, while a BC baseline trained from scratch on the blue-on-green data achieved only 0.5% success (1/200 episodes). Qualitatively, the BC baseline would consistently move towards the blue object and occasionally pick it up and place it on top of the green object, but a full, stable stack was almost never achieved.

3.5.4 Robotics: Skill Mastery

Similarly to the Skill Generalization challenge discussed in Section 3.5.2, the Skill Mastery challenge consists in training a robotic arm to stack blocks of different shapes. However, the Skill Mastery allows the agent to train on data involving the object shapes used for evaluation, i.e. the *test* set in Skill Generalization becomes a part of the Skill Mastery *training* set. Thus,

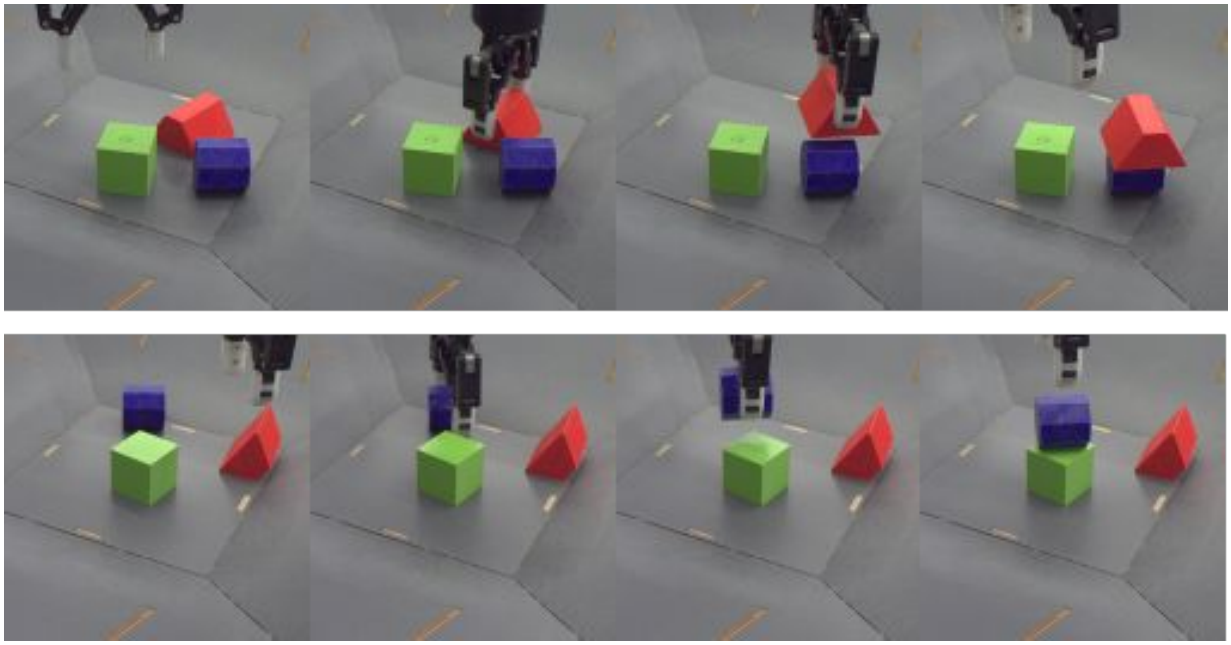

Figure 3.6: Comparing training/test task goal variations. Top: the standard "stack red on blue" task tested in the Skill Generalization benchmark. Bottom: the novel "stack blue on green" task demonstrating Gato's out of distribution adaptation to perceptual variations.

Table 3.3: Real robot Skill Mastery results. Gato is competitive with the filtered BC baseline.

Agent	Group 1	Group 2	Group 3	Group 4	Group 5	Average
Gato	58%	57.6%	**78.5%**	**89 %**	**95.1%**	**75.6%**
BC-IMP [52]	**75.6%**	**60.8%**	70.8%	87.8%	78.3%	74.6%

this challenge serves to measure Gato's performance on in-distribution tasks (possibly with initial conditions not seen in the training demonstrations). Our Skill Mastery results use an earlier version of the Gato architecture described in Appendix A.2, with no fine-tuning.

Table 3.3 compares the group-wise success percentage and the average success across object groups for Gato and the established BC-IMP baseline. Gato exceeds or closely matches BC-IMP's performance on all but one training triplet.

3.5.5 Simulated robotics ablations

I conducted a series of ablations in simulation to better understand the effect of diverse pretraining data in the robotics domain (see Figure 3.7). I included a variety of baselines

with a `364M` parameter size variant of Gato trained on different data splits of the pretraining set, including one trained on DM Control Suite data only. The DM Control-only agent is superior to the base Gato at zero-shot transfer and with a lot of fine-tuning data, suggesting that Gato may not be using the representations learned from the text-based datasets when adapting to robotics tasks. The same domain only agent performs the best overall, matching the CRR baseline at 1 fine-tuning episode and outperforming it with more data, suggesting that Gato at current scale can trade its generalization capacity for data-efficient and effective few-shot adaptation.

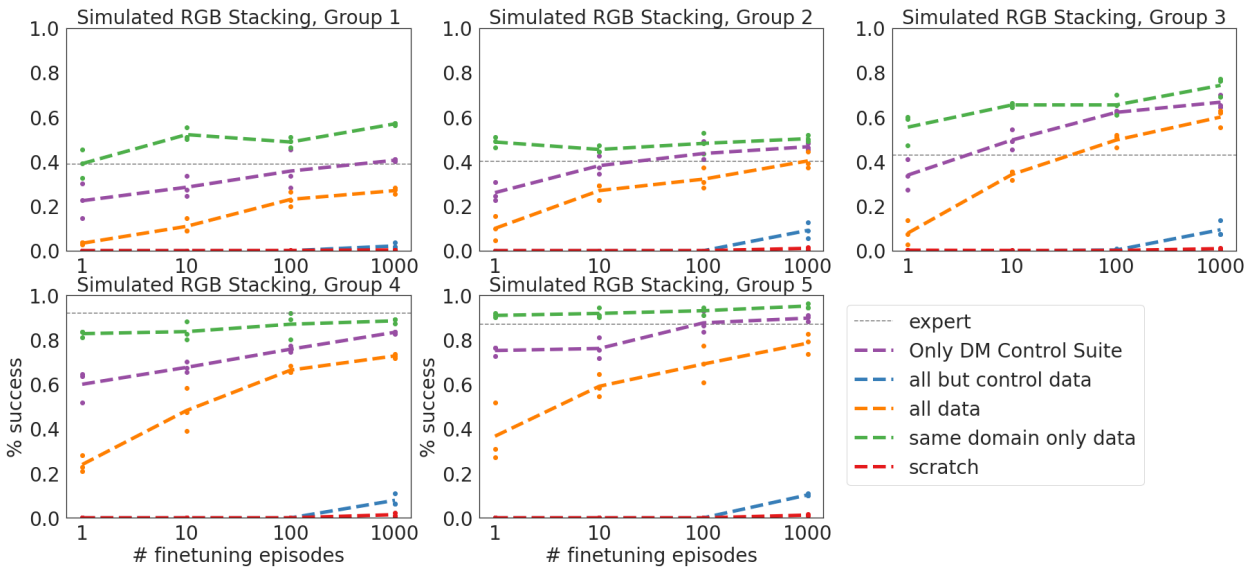

Figure 3.7: Few-shot performance of Gato for Skill Generalization in simulation. Each test set object is plotted separately. I ablate over different pretraining datasets.

3.6 Interpretability

We hypothesize that Gato's capacity for transfer is facilitated by inner representations that emerge during training and are recruited when the model is presented with relevant stimuli. In this section, I provide intuitive visualizations to argue how Gato's inner mechanisms facilitate acting intelligently in many domains. First, I computed attention maps as a proxy for saliency and superimpose them over Gato's visual observations to show that the agent attends to task-relevant objects (rather than simply memorizing actions correlating with specific observations). I also visualize clusters of inner embeddings by task to show that

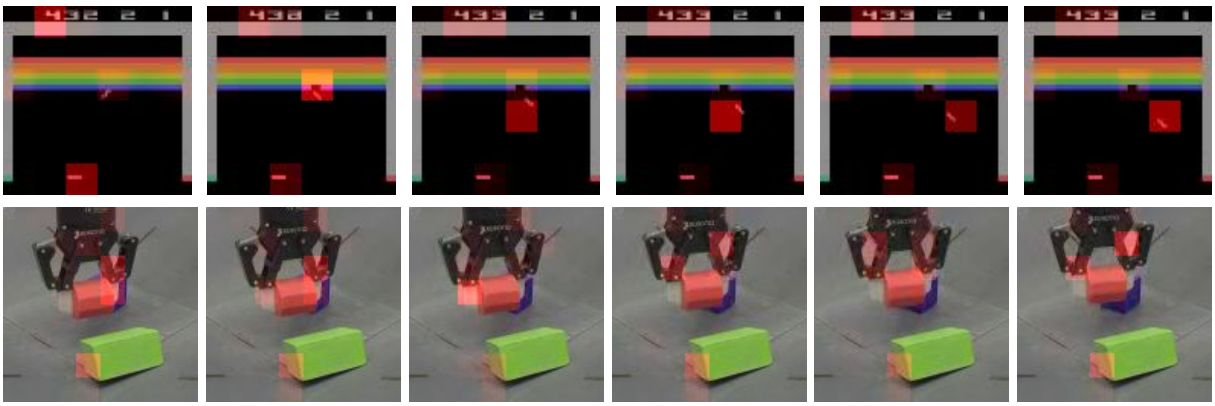

Figure 3.8: Attention maps. Time-lapse attention maps from selected heads at the first layer for Atari Breakout and RGB Stacking.

the embedding space reflects semantic similarity between domains. Although this analysis does not constitute rigorous proof of representations in the cognitive science sense [125], the techniques presented are a step in the direction of providing interpretable explanations for the emergent capabilities of generative AI.

3.6.1 Attention Analysis

I visualized the transformer attention weights over the image observations for various tasks, to gain a qualitative sense of how Gato attends to different regions of the image across tasks.

To render the transformer attention weights, I retrieved the cross-attention logits, a tensor with dimension (H, T, T) where H is the number of heads and T is the number of tokens in a sequence. The (h, i, j)th entry of this matrix can be interpreted as the amount that head h attends to token j from token i. Due to Gato's image tokenization scheme, there are multiple tokens per timestep. Therefore to render the attention for a particular timestep, I took the sub-matrix that corresponds to that timestep. We then applied a softmax over the rows of this matrix to normalize the relevant values. Because we are only interested in attention to the previous tokens, I excluded the diagonal by setting it to negative infinity before softmax.

To measure the importance of each patch, I averaged the attention weights over the corresponding column. Because Gato uses a causal transformer, the attention matrix is lower triangular, so the mean was only considered over the sub-column below the diagonal of the

matrix. This corresponds to the average attention paid to particular patch over a whole timestep.

Using this method, I found the attention maps at the first layer the transformer to be most interpretable, agreeing with the findings of Abnar and Zuidema [126]. Certain heads clearly track task-specific entities and regions of the image. Figures 3.8 and 3.9 shows the attention maps for manually selected heads at the first layer for several tasks.

3.6.2 Embedding Visualization

To understand how Gato encodes differently information per task, I visualized per-task embeddings.

I analysed 11 tasks. For each task, I randomly sample 100 episodes and tokenize each of them. Then, from each episode I take a subsequence of 128 tokens, compute their embeddings (at layer 12, which is half the total depth of the transformer layers) and average them over the sequence. The averaged embeddings for all tasks are used as input to PCA, which reduces their dimensionality to 50. Then, T-SNE is used to get the final 2D embeddings.

Figure 3.10 shows the final T-SNE embeddings plotted in 2D, colorized by task. Embeddings from the same tasks are clearly clustered together, and task clusters from the same domain and modality are also located close to each other. Even a held-out task (`cartpole.swingup`) is clustered correctly and lays next to another task from DM Control Suite Pixels.

3.7 Limitations

3.7.1 Data collection

Gato is a data-driven imitation learning approach. It therefore inherits many of the limitations of imitation learning, such as the inability to meaningfully exceed the performance of the demonstrator. Imitation is also highly bottlenecked by the quality of demonstrator trajectories. This is exacerbated by a limitation inherent to large-scale deep learning, where it can be difficult to gather enough high-quality data such that billions of parameters do not underfit

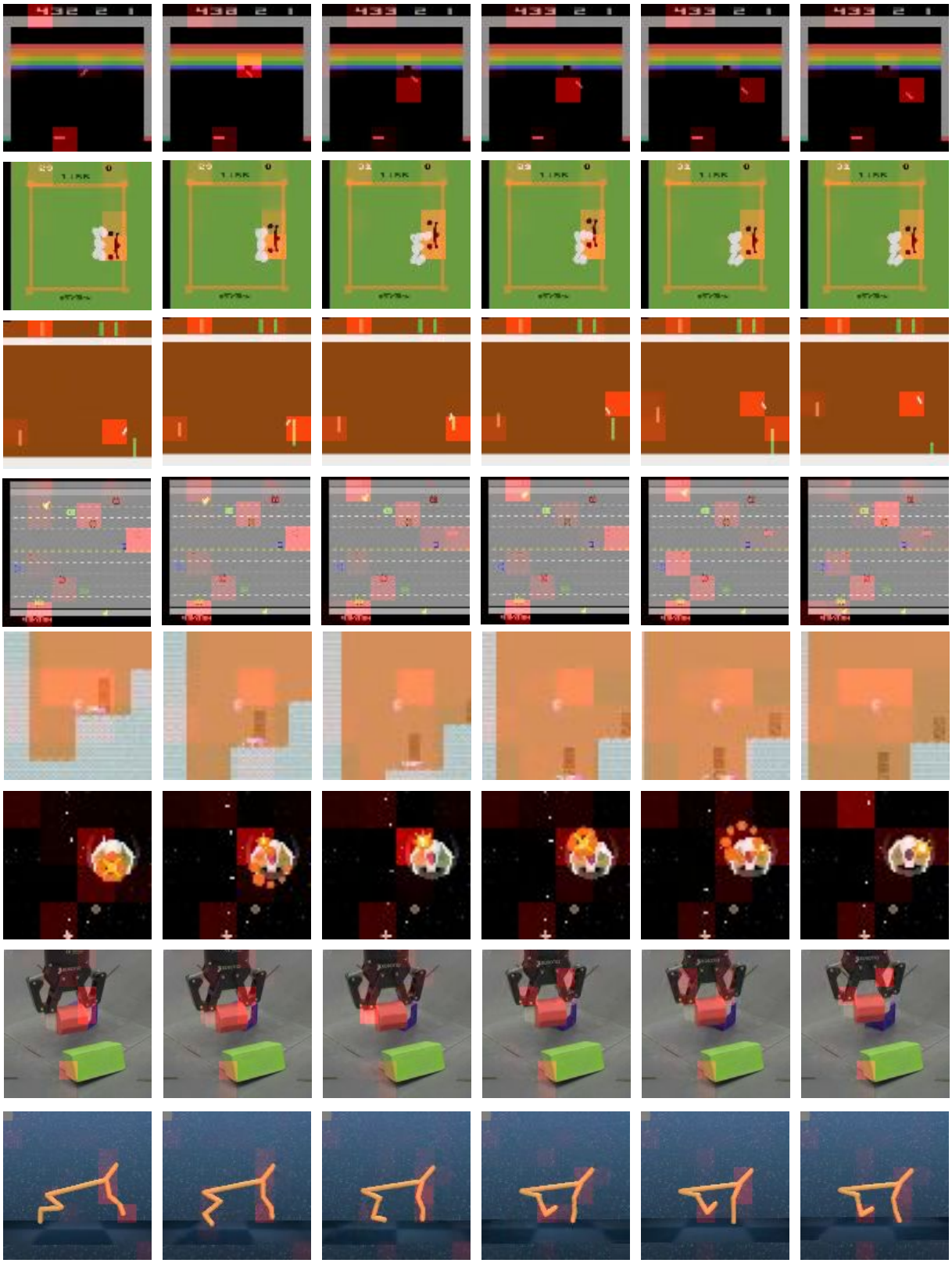

Figure 3.9: Attention maps. Time-lapse attention maps from selected heads at the first layer for Atari Breakout, Boxing, Pong, Freeway, Procgen CoinRun, Bossfight, RGB Stacking, and DM Control Suite Cheetah.

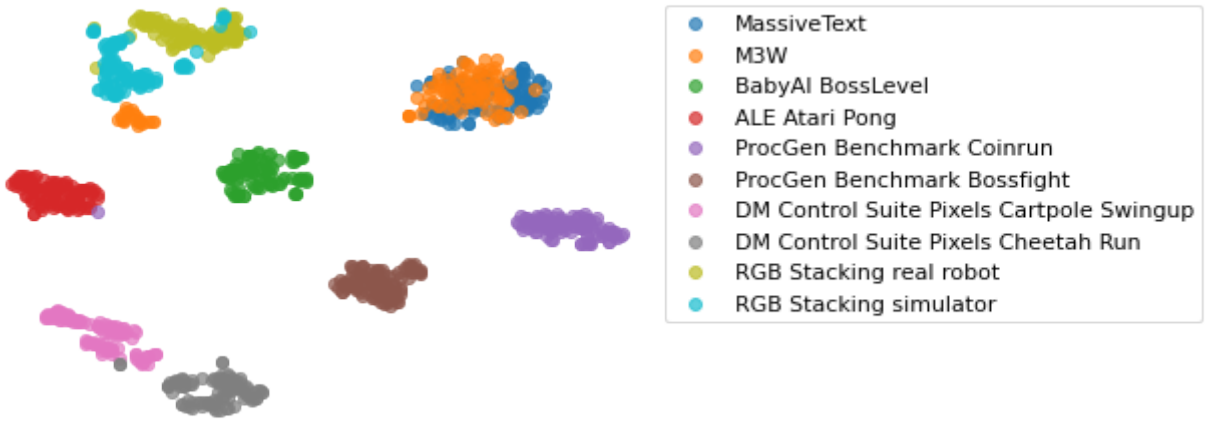

Figure 3.10: Embedding visualization. T-SNE visualization of embeddings from different tasks. A large part of the vision-language embeddings (M3W) overlaps with the language cluster (MassiveText). Other tasks involving actions fall in their own cluster.

to the training data. While natural language or image datasets are relatively easy to obtain from the web, a web-scale dataset for control tasks is not currently available. This may seem at first to be problematic, especially when scaling Gato to a higher number of parameters.

That being said, there has already been extensive investigation into this issue. Offline RL aims at leveraging existing control datasets, and its increasing popularity has already resulted in the availability of more diverse and larger datasets. Real-life data has also been already stored for ML research purposes; for example, data for training self-driving cars is acquired from recording human driver data. Finally, while Gato uses data consisting of both observations and corresponding actions, the possibility of using large scale observation-only data to enhance agents has been already studied [127]. Thanks to online video sharing and streaming platforms such as Youtube and Twitch, observation-only datasets are not significantly more difficult to collect than natural language datasets, motivating a future research direction to extend Gato to learn from web data.

While the previous paragraph focuses on alleviating drawbacks of data collection from RL agents, it is important to note that this approach presents a different set of tradeoffs compared to scraping web data and can be actually more practical in some situations. Once the simulation is set up and near SOTA agent trained, it can be used to generate massive amounts of high quality data – in contrast to web data, which is notorious for its low quality. A subsequent work to Gato investigated the automation of this data flywheel through interactive

generative *environments* by training a "world foundation model" [128].

Acquiring suitable, high quality data in diverse modalities is an important area of research. While this is generally true for machine learning research, the momentum behind foundation model-style training is rapidly expanding the field's investment in this area.

3.7.2 Prompt and short context

Gato is prompted with an expert demonstration, which aids the agent to output actions corresponding to the given task. This is particularly useful since there is otherwise no task identifier available to the agent (that is in contrast to many multi-task RL settings). Gato infers the relevant task from the observations and actions in the prompt.

However, the context length of the agent is limited to 1024 tokens, which may translate to only a few environment timesteps in total. This is especially the case for environments with image observations, where depending on the resolution each observation can result in more than one hundred tokens each. Hence for certain environments, only a short chunk of a demonstration episode fits in the transformer memory.

Due to this limited prompt context, preliminary experiments with different prompt structures resulted in very similar performance. Similarly, early evaluations of the model using prompt-based in-context learning on new environments did not show a significant performance improvement compared to prompt-less evaluation in the same setting.

Context-length is therefore a current limitation of the architecture, mainly due to the quadratic scaling of self-attention. Recently proposed innovations enable a longer context at greater efficiency [129], which could potentially improve the agent's performance.

3.8 Broader Impact

Although generalist agents are still only an emerging area of research, their potential impact on society calls for a thorough interdisciplinary analysis of their risks and benefits. For the sake of transparency, I documented the intended use cases of Gato in the model card included in A.1. However, the tools for mitigating harms of generalist agents are relatively underdeveloped, and require further research before these agents are deployed.

Since the generalist agent can act as a vision-language model, it inherits similar concerns as discussed in [61], [69], [115], [130]. In addition, generalist agents can take actions in the the physical world; posing new challenges that may require novel mitigation strategies. For example, physical embodiment could lead to users anthropomorphizing the agent, leading to misplaced trust in the case of a malfunctioning system, or be exploitable by bad actors. Additionally, while cross-domain knowledge transfer is often a goal in ML research, it could create unexpected and undesired outcomes if certain behaviors (e.g. arcade game fighting) are transferred to the wrong context. The ethics and safety considerations of knowledge transfer may require substantial new research as generalist systems advance.

Technical AGI safety [131] may also become more challenging when considering generalist agents that operate in many embodiments. For this reason, preference learning, uncertainty modeling and value alignment [132] are especially important for the design of human-compatible generalist agents. It may be possible to extend some of the value alignment approaches for language [133], [134] to generalist agents. However, even as technical solutions are developed for value alignment, generalist systems could still have negative societal impacts even with the intervention of well-intentioned designers, due to unforeseen circumstances or limited oversight [135]. This limitation underscores the need for a careful design and a deployment process that incorporates multiple disciplines and viewpoints.

Although still at the proof-of-concept stage, the recent progress in generalist models suggests that safety researchers, ethicists, and most importantly, the general public, should consider their risks and benefits. Gato is not currently deployed to any users, and so we anticipate no immediate societal impact. However, given their potential impact, generalist models should be developed thoughtfully and deployed in a way that promotes the health and vitality of humanity.

3.9 Conclusion

Although this chapter focused on imitation as a method, the other themes of the thesis present themselves as I consider the wider implications of building a generalist agent. Gato's generality opens up the possibility for unexpected or undesired behaviors to emerge. The impact of

this emergence, considered in 3.8, foreshadows the ethical concerns of generative models. In Chapter 5, I will articulate these concerns from a philosophical perspective, focusing on two major challenges which emerge from training large generative models: algorithmic injustice and misinformation.

Imitation of the successful strategies present in training data is the explicit goal of Gato's optimization objective. However, there is an overarching goal that Gato learns to *generalize* skills from the data distribution that transfer to unseen settings – thereby imitating the human capacity for rapid adaptation. While generalization and adaptation were sought after in this chapter, these properties are not necessarily linked to the technique or goal of imitation. The design of Gato tried to achieve a capability for generalized and adaptable imitation through learning from high diverse and general data. However, imitation learning from a narrow set of demonstrations should not be expected to exhibit generalization properties.

Does Gato really live up to the aspiration of a "generalist" agent? Certainly, it implementation of multimodality was a novel contribution at the time, enabling it to operate in a diverse and broad set of environments, and it exhibit a limited kind of generalization achieved through fine-tuning. However, we should be wary of overinflated claims of generality or generalization, or wild ambitions for generality. The narratives that motivate building all-powerful generalist technologies carry a concerning impulse for consolidation of power behind a singular method, which is also epistemologically detrimental in the long term [65]. Furthermore, Section 3.8 gestured at the ethical concerns that arise from general-purpose technology: its dual-use nature opens potential for misuse and an excuse that model developers can hide behind to shirk their responsibilities. The desirability of generalist agents is then called into question, a theme which will recur in Chapters 5 and 6.

In the next chapter, imitation will be the goal, but not the method. Motivated by the ethical harms of identity-based algorithmic bias, I propose to imitate the human representation of social identity. While Gato could model the environment to produce actions which imitated its demonstrators, Chapter 4 focuses on the modelling of human minds and expectations to inform the fluid and contextual nature of identity categorization.

Chapter 4

Representational Challenges in Human Social Identity

The previous chapter foreshadowed a concern that the identity-driven bias of humans would be imitated by large generative models, following an existing trend of algorithmic injustice. While technical AI fairness methods attempt to mitigate these biases, they depend upon a conception of human social identity as essential attributes that are discrete and static. In stark contrast, strands of thought within critical theory present a conception of identity as malleable and constructed entirely through interaction. This more fluid conception of identity is thought to be critical to challenging the oppressive normativity and power differentials between identity groups. Motivated by these ethical concerns, I now ask what it would mean to imitate a human-like understanding of identity. We now use imitation as a goal, not a method, by assessing existing AI fairness solutions against our theoretical ideal of identity. We additionally suggest alternative configurations of systems which may be better suited to the imitation and comprehension of human identity.

4.1 Introduction

Turning from imitation learning as a purely technical method, I now investigate the case of imitation as a goal. Rather than a means to an end, imitation is now the end itself: the overarching design goal, or the evaluation criterion. This may be necessary when the

imitated quality is not an essential attribute circumscribed by a training dataset, but socially contextual, dynamic property. Imitation of this kind is increasingly relevant as human-AI interaction systems become commonplace. While quantitative evaluation often expects a fixed, computable metric, many pertinent properties of a socially interactive system are contextually dependent on historical and cultural considerations, or philosophical definitions. When assessing imitation in an AI system, the deviation from a fixed goal may occur not due to any deficiency of the imitator, but because of subjectivities in the expectation imposed by human evaluators. Thus, evaluating imitation of social behavior requires a philosophical deepening of our analysis, beyond the offerings of mathematics and computer science.

The algorithmic representation of human social identity exemplifies the challenges of imitating social cognition, requiring a complex, multi-layered multi-agent model of nuance and context. It is a catalyst for algorithmic injustice, where the historical bias against marginalized identity groups is systematically automated and amplified. However, the technical solutions for correcting these injustices largely depend on insufficient ontologies of identity categorization. As this chapter will show, the default paradigm of identity in machine learning flattens these categories into discrete, static buckets which are held as the essential ground truth for algorithmic decision-making. Thus AI systems construct representations of humans which are composed of incomplete abstractions that cannot model the fluid character of social identity – a flawed imitation.

A useful concept of "identity" eludes machine learning practitioners, who are often caught between simplifying abstractions and stifling complexity. Inherited from a post-Cartesian empiricism that remains enshrined in much of scientific research, the default paradigm in AI research holds certain platitudes to be true about human identity: that it is composed of fixed, essential attributes. The basic tenets of statistical machine learning require either pre-existing categories of labeled data (in supervised learning) or seek to create such partitions within its training datasets (in unsupervised learning). In translating identity to data, arbitrary facets are sorted into *discrete*, often mutually exclusive categories. Such models tacitly assume that identity is fixed over contexts: *static* in its totality. They hold that identity is *essential*—that there is some intrinsic, self-encompassing, ground-truth quality to it. These conceptions neaten the problem of representation mathematically, but fall short in faithfully reflecting

identity as it operates interpersonally. Various components of identity are flattened when a person is represented by a system.

Strands of thought in critical theory and interrelated fields propose a far more dynamic understanding of identity than the simplifications made within AI research. In this chapter, I draw upon this theory for a critical analysis of the default representative paradigm of identity in machine learning. Among humans, identity encapsulates a system of social relations which are constantly in flux. It is announced and performed, ritualized and reinscribed, imposed by others or reclaimed. Identity emerges socially, and is reinforced through its usages: for drawing connections or distinctions between in-group and out-group, for dominance or for subversion, for recognition of the individual. A firm definition of identity is evasive, but theorizing about it as a *system of relational processes* is key. Incorporating these concepts into machine learning frameworks is necessary not only in terms of fidelity to reality, but also in building systems that do not freeze existing norms. In Section 4.2, I describe a theory of identity as an "autopoietic system," evolving processes of construction and function. A deeper substantiation of this theory which draws from fields of critical inquiry can be found in the original publication of materials included in this chapter, Lu, Kay, and McKee [136].

In Section 4.3, I critique existing AI fairness approaches that fail to incorporate this ontology by perpetuating a discrete, static, and essential notion of identity. I emphasize the risks that ensue from this type of abstraction, using the following three dichotomies to characterize identity and structure our analysis: discrete vs. continuous, static vs. contextual, and essential vs. co-constructed.

Finally, Section 4.4 imagines what an alternative framework that incorporates these qualities looks like. I outline two technical approaches to model design involving multilevel optimization and relational learning, then sketch an imaginary of what better systems could do. What possibilities are available to us with models that are capable of representing a fluid system of human social identity; what biases are circumvented; what avenues of feedback emerge afresh between human and machine?

Setting the right goal for the computational representation of identity – to better imitate what is observed in society – has far-reaching ethical consequences. By adjusting our collective grasp on what "identity" consists of as a field, and relinquishing poor imitations in favor of

more precise ones, I hope to mitigate harms caused by fundamental misconceptions within the frameworks that underpin the technologies we deploy. The effects of AI reverberate through society en masse; it is up to us what they will look like.

4.2 Identity as Autopoiesis

What do we mean when we talk about "identity"? Do we conceptualize it as a fixed substance interpolated through the world, or rather a discursive narrative we tell about ourselves and others? Identity has emerged as a shorthand for a cache of attributes to apply to people (race, class, gender, etcetera), but is best understood as an action: to identify. You identify *with*, or disidentify, using relational terms to situate yourself among the categorized crowd. Reconstrued as an action, we can then ask: identification with *what*? Identification operates in the realm of the imaginary, as a process of differentiation that involves superimposing layers of fantasy over the identified subject. Your fantasy, their fantasy, your fantasy of their fantasy, their fantasy of yours – recursive layers of interpersonal cognitive modelling. This is a non-deterministic, context-sensitive happening which nevertheless sediments into a sorting pattern, a complex classification, a "necessary negotiation between detail and abstraction" [137]. Our collective categorizations become a nascent topography, invisible and in flux, yet still with potent ramifications.

It is easy to suppose that underneath these fantasy layers there is some substance to which they are hooked, that the *what* of identification can be unearthed. But this *what* is indeterminate in the same way the process of identifying is: there is no fixed quality of a person that is constant across *all* identifications made by *all* people across *all* time. What remains in dialogue is the overlap of imaginaries within identification processes. On a collective scale, these coagulate into the social norms used for identifying, which Judith Butler calls "cultural intelligibility" in *Gender Trouble* [138]. What we recognise as identity's substance is itself a product of unfurling fantasies of what identity might consist of. The recursion is readily apparent when any perceived substance of identity is placed under close scrutiny. While identity's substance is often illusory, its downstream impact is undeniably material. Identity is integral to the human experience, philosophically and psychologically.

Its relevance to politics rests in its irreducible existence as a vector of power: how it mediates relationships between subjects. All perceived differences are capable of generating exclusion and power differentials, substantiated or otherwise.

In 1972, biologists Humberto Maturana and Francisco Varela introduced the term "autopoiesis" to describe a network of processes that is capable of reproducing and maintaining itself, in order to capture the self-contained chemistry of living cells [139]. Such a system would continuously regenerate its own components and organization, realizing the very processes which produced them. By conceptualizing identity as an autopoietic system comprising its *construction* and *function*, it is possible to foreground the self-reinforcing circularity and perpetual motion of its processes. *Construction*, therefore, refers to the processes by which identity is formed, discursively and psychically, throughout the mind that "owns" the identity as well as the others who ascribe it. *Function* refers to the processes through which identity is used, internally and interpersonally. Each set of processes informs the other, constituting the other's parts; the two can be co-located within the same social interaction. This abstraction allows for the contradictions inherent to defining identity. Feminist and queer theory disagree over how the body figures into identity, but this ontology makes no claim to locate from where precisely the construction or function is derived. Instead, it describes the processes by which the concept takes form.

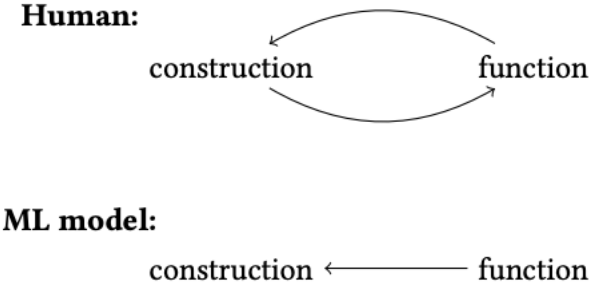

Figure 4.1: Diagrams for the identity processes of humans (bidirectional) vs. machines (unidirectional).

Visualizing identity as an autopoietic system of interaction reveals how it is capable of evolving. As it is constituted of feedback loops, perturbing any norm around either how identity functions or is constructed can cascade across the entire network. Social drift is possible. The machine learning abstraction of identity, however, is not that of an autopoietic

system, but rather a unidirectional one: how human identity is construed is derived purely from its utility to the algorithm (Fig. 4.1). This immobilizes the processes by which identity evolves in society as AI becomes a ubiquitous interlocutor in everyday life. Every existing category is reified. Without critically reframing identity in the field, hierarchies of power will calcify in the same way.

4.3 The Limitations of AI Fairness

In the humanities and social sciences, identity is a socially constructed, circularly reinforced system of formation and function. In comparison, machine learning systems enforce a unidirectional relation: categories are formed and fixed according to their utility to the model. AI fairness researchers have presented well-intentioned technical solutions to mitigate bias and enforce equality for protected groups. However, their abstraction of identity forms a precarious foundation for fairer systems. In this section I critique the ontological assumptions around identity in AI fairness by providing examples violating our theory of relational identity.

4.3.1 Discrete vs. continuous

Many AI fairness approaches assume that identity is *discrete*, using binary gender [140] or mutually exclusive racial categories [141]. This opposes a *continuous* representation of identity, which can be imagined as real-valued numbers. Even in one system using continuous inputs, the predicted attribute (rate of violent crime) is binarized to 0 (30th percentile of crime rate) and 1 (70th percentile of crime rate), to collapse the problem to classification [142]. Discretization means erasure when identity operates in a complex continuous space. Gender is not a binary, but rather a shifting discursive act. Multiracial people are not a single category, and lumping them together is harmful negligence. Furthermore people in the same racial "category" often receive different levels of discrimination due to colorism. When it comes to sexuality, the Kinsey scale reflects sexual orientation as a spectrum, rather than a binary of hetero- and homosexual (further enriched by multi-dimensional systems such as the Klein Grid and the Storms Scale [143]).

This lack of gradation often occurs because of the limitations of the chosen model, such

as when classification (which outputs discrete categories) is chosen for prediction instead of regression (which outputs real-valued predictions). In the case of the racial categories of Ionescu's study of race-based homogamy in [144], race is reduced to black and white because the agent-based model utilized could only handle binary features. Even the Gender Shades study, which analyzed colorism in gender classification systems, chooses discrete categories for labelling skin tone rather than a continuous variable [145]. The decision to discretize identity can be made anytime during system design, including during dataset collection and curation, before model constraints come into play. The ProPublica COMPAS dataset, created to analyze racial bias in recidivism predictions, divides race into six categories: black, white, Hispanic, Asian, Native American and "Other" (a category describing 343 out of 80,000 defendants) [141].

An argument for discretization is: people often make categorical judgments about social identity, so why should machines differ? Indeed, multiracial people are often treated as a single race at both an interpersonal and systemic level, called *monoracial normativity* by a 2019 study illustrating the phenomenon [146]. But human beliefs are ultimately malleable, while the categories programmed into machine learning systems are not. The risk of discrete categories emerges when machine inference informs action; after all, models are deployed so they can be used. A low-dimensional discrete prediction can only produce simplified actions that may not be proportional in context. In the case of binarized violent crime risk rates, if predictions decide whether extra police are deployed to a neighborhood, those on the precipice of the threshold will be particularly ill-served and experience a higher rate of false arrests and police violence. Deploying any predictive policing heuristic forms a feedback loop that reinforces carceral injustice, but this illustrates additional harms caused by discrete classification.

Using real-valued variables solves some of these problems by allowing gradations of difference. However, they do not fully satisfy the autopoietic model of identity. Continuous values do not capture the dynamism of identity. Even placed on a continuum, our skin color does not determine "how much" we identify with a racial category. Upbringing and cultural context informs our relationship with race and its perception (consider adoptees, immigrants, multiracial people). We code-switch, linguistically and behaviorally, based on

whether we are at work, at home, among friends, walking down the street. Indeed, this phenomenon is present for many facets of identity. What equation can possibly capture the social complexities of "how Black" someone is, or "how much of a woman"? The answers to these questions, insofar as they can be answered, differ across contexts.

4.3.2 Static vs. contextual

A *static* identity is immutable across time and unchanged by other variables in the data, observable or otherwise. In contrast, understanding identity as *contextual* recognizes how it changes across time and space. To contrast the continuous conception of identity with the contextual, consider a fictional example. Alice is labeled as "heterosexual", but actually identifies as "mostly straight": this continuous trait is discretized in most classifications. Meanwhile, Alice's colleague Bob records his identity as heterosexual on a workplace survey while seeking out only same-sex relationships on a dating app, because he prefers not to disclose his sexuality at the office. Now, imagine Bob's employer-sponsored insurance company deploys a ML system that ingests a massive amount of anonymized employer-gathered data as well as an employee's personal web browser history. Maybe the designers of the system want to detect the sexual orientation of employees because this is a protected characteristic, and they want to ensure fair outcomes for people like Bob. Are there frameworks in AI fairness that account for Bob's identity changing in context?

While some tools in machine learning handle context and mutability, identity labels are typically made static in AI fairness. In the influential frameworks of individual and group fairness, and offshoots such as subgroup and intersectional fairness, protected characteristics are considered static for all individuals. As a solution, counterfactual fairness has been proposed for incorporating causal context into fairness analysis. In this type of analysis, counterfactual reasoning can infer the influence of observable factors on other variables in the dataset. By constructing a directed acyclic graph of latent variables that influence observable variables, the inference process identifies causes via intervention: the substitution of different values for the latent variables [147].

However, there are numerous criticisms of how counterfactual fairness fails to accurately reflect identity dynamics. Kohler-Hausman argues that using the causal inference framework

to detect racism reduces race entirely to phenotype, neglecting the interplay of internal experience and societal sculpting [148]. Kasirzadeh et al. [149] offers a similar critique and advises systems designers to consider the semantic and ontological origins of identity categories in causal graphs. In [150], Hu recognizes the ontological problem of representing social categories as nodes on the causal graph, instead of regions in a dense, ever-shifting web of relations. These are corollaries to our theory of identity as an autopoetic system. We launch a complementary critique by acknowledging contextual identity. Counterfactual fairness does not allow identity to vary contextually, but instead compares individuals with varying sets of characteristics as counterfactual evidence for each other's outcomes. In this approach, interventions are made by generating fictional individuals with counterfactual identities. However, in a model of counterfactual fairness, the *same* individual would display different identity characteristics under different circumstances. This subtle difference has implications for fair outcomes. In fact, Kusner emphasizes in the supplemental of *Counterfactual Fairness* that the influence of protected attributes on decisions relies upon interventions *across* individuals, not *within* the same individual.

Beyond this, handling context does not complete a system's model view of identity. Context-sensitivity alone permits labels that are mutable but deterministic, in which a fixed set of circumstances can change the contextual variable. But a system that does not constantly revise its cycle of construction and function fails to fulfill our theory of identity processes. Epistemological questions are left unanswered. Who determines what entities and signifiers are contained in a dataset? How are the signs that comprise a model's conception of identity assembled? What are the dangers of assuming their ground truth?

4.3.3 Essential vs. co-constructed

Machine learning systems tacitly assume that the signifiers within its data reflect inherent properties of the entities represented. Predictions, inferences, and calculations operate within the rigid semiotics set by the design of the dataset and the model. When they include identity labels, the system interprets the classification of people as a knowable, objective, *essential* truth. This contrasts the theory of *co-construction*, which frames identity as an ongoing interaction, generated indeterminately. While discrete simplification begs the question, "Can

one's identity exist on a spectrum?", and static simplification begs, "Can one's identity change over time and space?", the essential simplification escalates: "What is actually represented by one's categorical signifiers, and from whose point of view?" Essentialism assumes identity is an innate property capturing the subject's "essence."

We can connect essentialism in machine learning to a colonial scientific tradition running through inequitable medical treatment [151], eugenics [152] and the reinforcement of colonial power and norms [21]. In *Truth from the machine*, Keyes et al. present a case study of how the assumption of inborn causes for autism and homosexuality led to faulty conclusions in data science-driven studies [153]. Sociotechnical systems contain categories within their structure, and thus any system attempting to enact fairness along axes of identity necessarily reinforces their boundaries. Such a system may introduce harms to invisibly marginalized groups (though invisibility can even be desirable). Such systems also have long-term implications, by perpetuating a static set of standards while social norms drift. Debiasing machine learning systems with respect to protected characteristics is a significant goal for AI fairness research, but if protected categories are drawn along essential lines, these systems risk reinforcing systemic inequality.

Recent works interrogate the epistemic basis of identity representations in AI. Hanna, Denton, et al. [154] set out guidelines for applying critical race theory to understand algorithmic categorization of race and ethnicity. Denton additionally describes a genealogy of AI in order to confront the histories and norms embedded in datasets [155]. The practice of "studying up" in machine learning reverses essential assumptions by turning the lens of machine predictions on the social norms and cultural context of those holding power [156]. In a similar vein to this work, Hancox-Li et al. critiqued the limitations of feature importance methods and suggests methodologies from feminist epistemology to address them [157]. In their writing, context sensitivity and interactive ways of knowing agree with our theory of identity as ongoing relational processes.

The co-constructivist stance holds semantic categories as fluid and subjective. For any given identity category, people have their own adaptable notion of its semantic meaning, regardless of their relation to it. It is therefore impossible—or at least ill-advised—to build a machine learning system that discovers, classifies or infers "objective" identity labels, or

draws "objectively correct" conclusions from identity data. Prior works either elide this point or leverage it to critique essentialist paradigms in science and technology without proposing alternatives. In contrast, the next section takes a positive position. Humans make subjective classifications all the time. We constantly update our abstractions; we hold contradictory viewpoints; we imagine the experiences and preferences of others. Ultimately, our conceptions of identity are fluid and susceptible to change, play, and subversion. Is it possible to change conceptions of identity within machine learning models to imitate us?

4.4 Alternative System Configurations

While the field of AI fairness offers a growing number of technical solutions, these often operate under the same assumptions as the models they criticize. The majority of solutions do not allow the circular, iterative nature of identity; rather, they continue to assume it is composed of fixed, essential attributes. They immobilize identity concepts and do not permit the possibility of drift or subversion. Broader critiques of machine learning fundamentals are more aligned with this paper's theory of identity as interaction, but they do not venture positive solutions, and some even claim the discipline itself is fundamentally at odds with the ambiguity of human behavior [158]. I acknowledge with full modesty that machine learning systems have epistemic limits when it comes to understanding of identity. However, within these limits exists a rich space of system configurations that are underexplored in the applications of AI fairness. Conceptualizing identity as autopoiesis, comprising processes of *construction* and *function*, also permits new imaginaries about how interaction *between* identity systems and machine learning systems can play out: intra-system interplay. I offer two provocations to practitioners contending with AI fairness and identity:

1. Within machine learning systems, how can we close the loop from identity *construction* to its *function*? How can machines form internal concepts of identity that allow: mutability, iteration, social drift?

2. What interplay is possible when such machine learning systems are integrated *into* identity systems? What new forms of co-construction, destabilization, and subversion

are possible with the machine as interlocutor?

We hope these open questions stimulate new avenues of research, and map out possible directions by offering concrete technical frameworks and loose theoretical imaginaries. To close the loop within machine conceptions of identity, I sketch two complementary approaches: multilevel optimization and relational learning. To imagine new forms of identity co-construction between human and machine, I finish by illustrating scenarios of intra-system interplay, when machines *can* conceive of fluctuating identity.

4.4.1 Autopoiesis as multilevel optimization

Society collectively constructs a concept of identity to suit certain functions, but individuals also define necessary functions based on their identity. Can we capture this cyclic interplay of construction and function with a computational model? A possible tool for describing our model of identity is *bilevel optimization*, in which one model is optimized with respect to the optimum of another:

$$y^*(x) = \arg\min_{y \in Y} f(x, y)$$
$$x^* = \arg\min_{x \in X} F(x, y^*(x))$$

The family of bilevel optimization encompasses two popular machine learning frameworks: generative adversarial networks (GANs) and actor-critic methods in reinforcement learning [159]. We separately consider these frameworks to understand if they suit the case of learning identity, then sketch the schematics of an identity learning system inspired by these techniques.

In the unsupervised GAN setting, the goal is to learn how to generate likely samples from the distribution of the input dataset. A generator network produces candidate samples to fool the discriminator, while a discriminator network classifies samples as real or generated. An obvious application to identity is to train the discriminator to classify (or regress) samples to an identity label, and optimize the generator to produce samples that defy categorization. These adversarial identity examples could qualitatively illustrate characteristics that resist implicit norms. This approach also expresses the contextual, fluid property of our model of autopoietic

identity. Identity boundaries shift over time as the generator and the discriminator interact. However, this setting requires supervision in the form of identity labels for people represented in the training data. Thus, this approach fails to overturn the default representation of identity in machine learning. In the final artifact of training, a contextual understanding of identity is not guaranteed. The features that constitute an identity category are essentially and unidirectionally determined by the dataset. Perhaps a different extension of GANs can more closely fit our model of identity, but I leave that exploration for future work.

In the actor-critic setting, the goal is to train an agent to output a sequence of actions that maximizes reward over time in its environment. A policy network produces actions that maximize the value function, which is learned by the critic function. Here, value is defined as the discounted sum of expected reward of a state/action pair (contingent on the policy, since it is the source of future actions). In contrast to the adversarial setting, the two objectives are not opposed, but their circular dependence can make optimization difficult in practice [160]. Imagine a Markov Decision Process (MDP) where the states are observable characteristics of a person over time and the observed reactions of others, while actions represent the person's identity labels in that state. In this setting, policy learning is analogous to the individual's construction of their identity: what they choose to identify with over time, shaped by perceptual inputs, conferral of others, and individual preference. The critic represents the utility of the person's active identity in context. The actor-critic framework is not limited to discrete or continuous representations. The sequential nature of MDPs expresses the mutability of identity over time, which is missing from the GAN approach. Additionally, the actor and critic mutually construct features that constitute an identity from external and internal feedback.

What does the reward function represent in this scenario? One possible interpretation is that it is the individual's satisfaction with the identity label provided by the policy in the current context. Nonetheless, issues arise from this approach. Is satisfaction actively measured, or modeled? If humans are involved, how are they incorporated into the training loop? How will the system respond to identities not encountered during training? In an alternative configuration, the reward instead represents satisfaction within a specific task. Identification then becomes an auxiliary task that may or may not prove relevant, depending

on how the human's preferences interact with the environment. The policy associates identity actions based on their utility for the task at hand, according to the human's preferences, the responses of humans, or even the responses of agents. This proposal for actor-critic methods reiterates the importance of considering the potential uses and misuses of RL in AI fairness.

In contrast to many of the systems discussed above, this framework avoids the explicit priors that bias and immobilize identity. Nonetheless, it does not resolve epistemic issues in the dataset collection process, the choice of categorizations and representations, and the subjective positions within identity. In addition, it poses identity a function of the *environment*, which is ill-defined and may elude a useful operationalization. An environment enables the policy to learn identification must include both the preferences of the individual described and their surrounding social network. Consequently, in the next design sketch, I turn my focus to the relational nature of identity.

4.4.2 Relational and subjective learning

Hegemonic forces shape machine learning and are reflected in the systemic bias entrenched in standard datasets and benchmarks. For a case study of the power dynamics at play in the dataset collection process, see Miceli et al. [161]. As an alternative to Western universalism, Birhane proposes relational ethics as the guiding principle for more equitable sociotechnical systems [162], which center the interconnectedness of all entities and define personhood in terms of an individual's relationships to others. Relational schools of thought include Afrofeminist philosophy, the Zulu tradition of Ubuntu, and Eastern traditions such as Daoism. It also motivates an ecofeminist approach by respecting the interdependence of humanity and nature. [163] explores core concepts from Ubuntu to form a guiding framework for AI governance. However, relationality has not seen much adoption in AI fairness systems design.

The autopoietic model of identity is fundamentally relational: it situates the individual's identity within a shifting network of interaction. Each individual in a social network will perceive another's identity differently based on their individual experience. What would a relational approach to *learning* identity look like? One might begin with a subjective dataset of identifications. We can ask individuals in a community of interest to first describe their own identity, then describe the identities of the other participants, based on their subjective

internal logics. Such a dataset could train a wholly relational model of identity—subjective by construction, but reflective of the relations each individual brings to the collective identity system. We can condition a supervised classifier of identity on aspects of the beholder's self-identification (or relevant features of a latent identity embedding space, depending on the representation). In settings that also include data relevant to a resource distribution or allocation tasks, we can apply algorithmic fairness techniques to compute a set of subjectively fair outcomes, according to the different perceptions of identity and protected characteristics of individuals within the dataset. This operationalization of relations may be generally useful for capturing a network of semantics, where meaning is subjective and classification is contextual—from interpretations of the law and ethical codes, to biological taxonomies of organisms.

Of course, by definition this model will only generalize to the demographics of the pool of participants. As stressed by Bowker and Star [137], all standards reinforce and erase; there is no "view from nowhere." If certain groups are underrepresented with regards to the purpose and locale where the model will be deployed, then the quality of the conditional model will necessarily deteriorate. The model will not extend past the perceptions and biases of those groups which form its training data.

The format of the data is also a key consideration. Rather than the typical predetermined vector with constrained dimensions containing options for race, gender, etc. I propose recording identity as a freeform text field with an open prompt. Training a language model on such a dataset, if it were sufficiently large, could yield a rich, varied latent representation of identity, and a conditional model for identity generation. The externally perceived characteristics of individuals are open to customization based on the application of the system. Examples inspired by other fairness applications include facial photographs, resumes or job applications, healthcare records (noting that privacy should be carefully considered in any human dataset). As theorized by Stuart Hall [164], any observable sign of identity is "never a proper fit," always an incomplete snapshot of a shifting landscape. The proposed dataset would offer many snapshots of the landscape formed by its subjects, and the role of the algorithm would be to stitch those snapshots into a map of relational identity.

4.5 Conclusion

Theorizing about identity is mired with contradictions and indeterminacy. Rather than attempt to put forth a strict specification, I conceptualize identity as an autopoietic system. I understand it as processes of construction and function, that are cyclic and self-reinforcing yet pliable and amenable to subversion; this conception provides a foundation for critiquing existing paradigms in AI and imagining new possibilities. Common assumptions made in machine learning sever the bidirectional relations within this network by fully configuring identity according to utility, and not allowing the opposite causal flow. AI fairness techniques meant to ameliorate identity-based harms often invoke the same assumptions: that identity consists of discrete, static, and essential attributes. I argue these practices erase the elements that evolve identity concepts, preventing social drift, subversion, and the possibility of open-ended reinvention.

Identity may be part-situated in fantasy, but this does not denigrate its impact or validity. In Section 4.4, we offer two provocations, asking how we can close the loop *within* machine learning systems' conception of identity, and what futures are available when machines are playmates in identity systems rather than adversaries. I outline a high-level schematic for possible systems with multilevel optimization and relational learning, and sketch out a new imaginary for human-machine identity formation. Identity representation itself is an emergent property of human social interaction. Perhaps it, too, could emerge in machines.

As artificial intelligence becomes ubiquitous in our lives, it plays a burgeoning role in our identity relations. Given this context, I call for a fundamental re-imagining of how we configure our machines. It is up to us whether we make machines that calcify existing identities and concurrent power hierarchies, or machines that help us expand the dimensions of possibility. Will machine imitation of humans contribute to this expansion? In this chapter, I have explained how the traditional imitation of identity representation in AI is incomplete and suggested an alternative conceptual model. But the utility and ethical benefit of imitation is context-dependent, and more accurate reproduction of human behavior and cognition is not always desirable.

In fact, the unintended imitation of human qualities may emerge from the dynamic

interactivity of complex systems. Furthermore, although this chapter emphasized the fluid and part-fantastical nature of identity, it also emphasizes that identity-based power imbalances are material and of grave ethical consequence. The next chapter will argue why identity-based prejudice is harmful to our epistemic commons – our shared capacity for knowledge and understanding–and explain how it emerges in generative AI.

Chapter 5

Epistemic Injustice in Generative AI

While imitating some aspects of human behavior can be seen as a desirable goal, undesirable aspects of humanity can also emerge in large-scale machine learning models. This chapter focuses on the issue of *epistemic injustice* and investigate the emergent imitation of this injustice by large-scale generative models. It presents a systematic formulation of the ethical harms that can arise from the self-supervised imitation of foundation models. The theory of algorithmic epistemic injustice is illustrated by real-world examples that reveal how generative AI can produce or amplify misinformation, perpetuate representational harm, and create epistemic inequities, particularly in multilingual contexts. Examining epistemic injustice can also inform the development of epistemically just generative AI systems, through informing system design principles, and two approaches that leverage generative AI to foster a more equitable information ecosystem.

5.1 Introduction

In this chapter, I examine how generative AI methods, similar to the one used in Chapter 3, can result in unintended societal impact through the phenomenon of imitation as *emergence*. I define generative AI as the class of machine learning models trained on massive amounts of data, typically media such as text, images, audio or video, or in the case of Gato, observations and actions of an environment, in order to produce representative instances of such media [165]. There are many entangled factors in the societal impact of generative models, which

can be difficult to decompose. The technical parameters of a model's training and fine-tuning process, such as its data and architecture, interact with the sociocultural dynamics of the world outside of the machine learning lab. Generative models form semantic associations by learning probabilistic next token prediction. These associations inherit from the cognitive biases and prejudices held by humans. In the context of real-world interactions, behavior emerges that imitates the unjust circumstances that produced the model. Alternatively, deviations from this manifold of associations can also result in human users being misled in novel, unanticipated ways.

The emergent behaviors of generative AI has created an illusion that these systems are capable of "knowing" things. But what would it mean for an AI system to "know"? The act of knowing requires more than simply storing and retrieving information embedded in a neural network, which is roughly what these systems do. Knowing implies the ability to justify beliefs about the world, often through reflection upon lived experience. Generative AI systems cannot interact directly with the physical world the way humans do, and it is very difficult to argue that they have any lived experience to reflect upon. While in this chapter I will argue that the class of technology known as generative AI can perpetuate epistemic injustice, this does not mean that generative AI systems are "knowers" in the rigorous philosophical sense. However, as I will argue in this chapter, the development of generative AI, its usages, and the narratives perpetuated by the technology companies selling it are all potentially corrosive to the epistemics of humans. Generative AI does not need to have epistemic status equivalent to humans to be instrumental in epistemic injustice.

While algorithms have traditionally been leveraged to present and organize human-generated content, the advent of generative AI has started to fundamentally shift this paradigm. Generative AI models can now create content – spanning text, imagery, and beyond – imitating that of authors, journalists, painters, or photographers. The rapid advancement of generative AI, marked by accelerated software and hardware innovation and a proliferation of novel applications, has been accompanied by growing societal concerns and numerous instances of misuse [130], [166], [167]. These range from parroting harmful stereotypes and misconceptions about certain social groups [168], [169], confabulating facts

and distorting truth [170], and spreading misinformation and deepfakes [171], [172] [1]. Despite these escalating societal concerns and numerous instances of misuse, the discourse surrounding generative AI's rapid advancement lacks a philosophical account that coherently relates these epistemic concerns and explains how they constitute moral violations of a unifying principle.

To address this gap, I develop an account of *generative algorithmic epistemic injustice* by building upon an established philosophical understanding of *epistemic injustice*. Epistemic injustice emphasizes how identity-based prejudice within an information ecosystem not only unjustly hinders the expression of marginalized groups, but also significantly impairs the knowledge formation capabilities of all individuals. It does this by describing the harmful effects and ethical shortcomings inherent in knowledge production systems marked by hierarchical power imbalances rooted in identity.

While traditional discussions of epistemic injustice have primarily centered on interpersonal human interactions [174], [175], existing research on algorithmic epistemic injustice has largely been limited to epistemic injustices produced by decision-making and classification algorithms. However, I argue that the distinctive characteristics of generative AI give rise to novel forms of epistemic injustice that necessitate a dedicated analytical framework. To address this, I expand upon the established philosophical discourse on epistemic injustice and introduce an account of "generative algorithmic epistemic injustice," or simply "generative epistemic injustice," to characterize the variety of epistemic harms arising from generative AI systems from a philosophical standpoint.

In Section 5.2, I describe epistemic injustice as a social theory and argue for its ethical importance. Section 5.3 situates this chapter within the context of prior research on algorithmic epistemic injustice. Section 5.4 builds on the existing research on algorithmic epistemic injustice and identifies four distinct configurations of generative epistemic injustice: amplified and manipulative testimonial injustice, along with hermeneutical ignorance and access injustice. I illustrate these configurations through real-world exemplars of generative AI deployments. While the evidence of injustice in these systems is overwhelming, Section

[1] While a rigorous philosophical treatment of "truth" in the generative AI context is complex [173, p.9] and out of scope, my modest working definition is: statements that can be verified through adequate evidence, or a robust consensus between relevant social groups. The verification will differ based on the premises of the statement.

5.5 explores how we can design and use generative AI to foster epistemic justice. I propose strategies for resistance in the face of epistemic oppression caused by generative AI. For the original publication of this work, see Kay, Kasirzadeh, and Mohamed [176].

By unifying the epistemic injustices of AI misinformation, bias, representational harms, and power imbalances within the AI industry under a single four-dimensional theoretical framework, I identify commonalities in their mitigation strategies and build solidarity among affected groups. Although this phenomenon arose from an accidental imitation of humanity's undesirable qualities, the interventions for rising up against epistemic injustice could require both re-accessing our humanity and acting in entirely new ways.

5.2 Epistemic injustice

In 2013, the city council of Flint, Michigan decided to switch its water supply to the Flint River, notorious for its pollution from automotive manufacturing. This move swiftly provoked public outrage as residents reported tap water turning discolored and emitting a foul, sewage-like odor. Despite these immediate alarms, authorities repeatedly dismissed these complaints, perpetuating a longstanding pattern of environmental gaslighting. The situation escalated when an outbreak of Legionnaires' disease was revealed, previously concealed by political maneuvers to safeguard reputational interests. It was only after academic investigations uncovered alarmingly high levels of lead contamination that the gravity of Flint's water crisis gained national attention, unmasking a profound public health catastrophe. It is obvious in hindsight that the citizens of Flint were correct and the politicians and others in power committed egregious harm by not believing the residents' testimony [177].

Many have studied the tendency for those with more privilege to ignore and even oppress ordinary citizens with less privilege (socioeconomic, racial, gender or otherwise). The decolonial scholar Gayatri Spivak put forward the position that the elite's construction of an underclass, and their tendency to presume to speak on behalf of those they disenfranchise, means that the means for resistance to oppression are mediated through the oppressor. Thus according to Spivak, the "subaltern" – a group which includes women, the working class, the lower castes, citizens of "third world" countries, the colonized, and especially intersections

therein – cannot speak [178].

But Spivak's provocation begs the question: can the subaltern speak outside of the structures of the elite? Seeking alternative means of empowerment, Black feminist scholars devised systems of knowledge outside of white patriarchal domination. Patricia Hill Collins introduced a feminist epistemology that emphasizes the intellectual importance of wisdom gained through Black women's lived experiences, and the transmission of this wisdom through relationships, community, and solidarity [179]. For Collins, critical dialogue and resistance to threats of epistemic violence is necessary for assessing the claims of the powerful.

Later, Miranda Fricker brought these notions of oppression and silencing to the forefront of mainstream analytic philosophy by coining the term "epistemic injustice", referring to injustices related to knowledge and understanding [180]. However, Fricker's work has been criticized for not fully acknowledging the contributions of Black feminist scholars who had previously explored similar ideas [181]. These critiques highlight Fricker's oversight of intersectional aspects of oppression [182] and her failure to recognize women of color as agential knowledge creators. While acknowledging these limitations in Fricker's original account, her philosophical framework of epistemic injustice is theoretically expressive and influential in subsequent discussions of algorithmic oppression. Therefore, I use Fricker's account as the foundation for my philosophical examination of generative epistemic injustice.

"Epistemic injustices" includes scenarios where individuals or communities experience unjust discreditation of their knowledge and experiences due to underlying prejudices against their identity. Fricker emphasizes that these biases are directed towards groups with less social power, defined as the capacity to influence others' actions within social interactions and environments. This power can be either agential, exercised by individuals, or structural, embedded in cultural norms and material inequalities.

Fricker argues that epistemic injustices harm not only the oppressed but everyone. Sharing knowledge and experiences is a fundamental human right, essential for self-expression, establishing connections, and asserting needs. To deny this right based on identity is a discriminatory act. Knowledge acquisition is integral to human existence, shaping our understanding of the world, informing our interactions, and fostering a sense of purpose. Disregarding someone's testimony unjustly not only harms the individual but also deprives

society of valuable insights, obstructing collective knowledge growth. Those who unfairly disbelieve someone also do a disservice to others, blocking access to potentially valuable information. Moreover, recognizing and crediting someone's testimony is a valuable heuristic for anticipating and preventing harm.

Fricker distinguishes between two types of epistemic injustice: testimonial and hermeneutical. Testimonial injustice involves the unfair discrediting of someone's account due to prejudice against their identity, a recurring injustice experienced by the Flint residents in my opening example. Hermeneutical injustice, the second type, stems from a disconnect between personal experiences and societal understanding. "Hermeneutics" means the interpretation of knowledge. I will sometimes refer to our "shared hermeneutics" or "hermeneutical resources" as our shared cultural concepts for interpreting each other's experiences and for sharing knowledge.

Hermeneutical injustice occurs when a person's experiences are misunderstood or not recognized at all, due to the absence of appropriate concepts within our collective cultural knowledge. Shared knowledge is crucial for interpreting and relating to the experiences of others. However, the repository of collective understanding bears the imprints of dominant groups, leaving the experiences of marginalized communities underrepresented or distorted. This results in gaps and misinterpretations within our interpretive resources, leading to hermeneutical injustice. The transgender community, for instance, has frequently faced this form of injustice. In societies lacking widespread understanding of gender variance, fluidity, non-conformity, and the spectrum of body dysphoria/euphoria, the experiences of transgender individuals are often misinterpreted. This ignorance, as Fricker notes, leads to a lack of empathy and understanding. Beyond epistemic implications, the material consequences can be severe, obstructing access to essential social resources like healthcare, employment, and housing [183]. Both ignorance and misunderstanding contribute to hermeneutical injustice. [2]

To summarize, the distinction between testimonial and hermeneutical injustice lies in the assignment of credibility versus the availability of interpretative resources. In testimonial

[2]Medina [184] distinguished between "recognition deficits", in which a group is unseen or illegible, and "misrecognition", where a group is visible but misunderstood, subjected to false and distorted narratives. In a recognition deficit, the receiver of injustice is ignored or not recognized within the societal context. In misrecognition, the receiver of injustice is the subject of a statement which is false due to misrepresentation of their identity.

injustices, the speaker suffers injustice through the degradation of their credibility, a result of identity-based prejudice. Conversely, hermeneutical injustice relates to the absence of a lexicon to articulate the oppression they experience – leading to an inability to articulate an account altogether.

5.3 Related work on algorithmic epistemic injustice

The conventional account of epistemic injustice described in Section 5.2 only involves human actors. However, AI algorithms can also contribute to these injustices because they are epistemic technologies [185]: they consume, curate, and produce information, which is a precursor to knowledge. When algorithms make decisions, particularly in our bureaucratic systems, they enable businesses and governments to exert power. These organizations–often individual leaders, decision-makers or administrators–can use the outcomes of algorithmic decision-making to justify their own biases or make faulty inferences, thus accruing knowledge from AI systems that distorts their epistemics.

The emerging field focused on the intersection of epistemic injustice and AI algorithms has been named "algorithmic epistemic injustice," a term coined by Byrnes and Spear [186]. To situate this work within the broader landscape of algorithmic epistemic injustice, I begin by reviewing the existing literature on this topic. Several key themes emerge from this survey. Testimonial injustice can arise when algorithms are prioritized over human credibility, potentially amplifying existing societal biases. Additionally, hermeneutical injustices can occur when algorithms independently construct meanings and interpretive frameworks, often in automated setups without direct human oversight.

This body of research has a notable emphasis on classification and decision-making algorithms. Several studies exemplify this in different contexts. In child welfare systems, Glaberson [187] identify epistemic injustice through algorithms that disproportionately target Black communities and poor single mothers. These algorithmic testimonial injustices lead to wrongful mistrust, surveillance and over-policing committed by humans. The healthcare sector, as Pozzi [188] note, witnesses "automated hermeneutical appropriation" in opioid risk score predictions. Pozzi claims that the algorithm establishes meanings for concepts that

contribute to a patient's diagnosis, such as "addiction" or the experience of pain, without human intervention. The opacity of data science systems is highlighted by Symons and Alvarado [189], who examine the real case of a prisoner who was wrongfully denied parole by the COMPAS recidivism algorithm and continued to be detained, even after providing evidence for his case to human supervisors [190]. The authors argue that this lack of transparency facilitates epistemic injustice: the ignorance about a technology's inner workings complicated the possibility for contesting its unjust decisions. Hull [191] discusses how COMPAS and similar systems enable human carceral decision-makers to commit hermeneutical injustice through the algorithm's biased and stereotype-based classifications. Hull also points out the testimonial injustices inherent in physiognomic systems, which wrongly infer personal characteristics based on visual appearance, often linked to race, ethnicity, and gender. Building on these individual-focused analyses, Milano and Prunkl [192] emphasize the importance of our relationships to others for transmitting collective knowledge, which they call our shared epistemic infrastructures, and use these concepts to identify how algorithmic profiling can harm this infrastructure. This approach informs the subsequent discussion on access injustice in Section 5.4.4.

The domain of AI fairness research has started to intersect with broader concerns of social injustice [162], [193], [194]. There has also been a growing recognition of epistemic injustice concepts in relation to AI fairness. For instance, Edenberg and Wood [195] suggest that an epistemic lens offers a theoretical foundation for understanding the harms of algorithmic bias, which other prevalent frameworks might not adequately capture.

A notable difference between this cluster of prior work and ours is its primary focus on classification or decision-making systems. I build on this cluster to develop an account of epistemic injustice in relation to generative AI, which occupies a growing unique position in the landscape of epistemic injustice due to its capacity to produce a seemingly authentic output – a convincing imitation. While classification AI is set to delineate true categories and false ones, generative AI offers statements that play the role of testimonies, explanations, and interpretations. However, the quality and veracity of these outputs varies, posing risks of epistemic contamination.

To the best of my knowledge, the only work with a focus on the epistemic injustice of

Table 5.1: Summary of the four configurations of generative epistemic injustice and their defining examples. I also summarize the corresponding interventions for achieving epistemic justice proposed in Section 5.5.

	Example	**Intervention**
Amplified testimonial	Parroting misinformation	Identify testimonial injustice
Manipulative testimonial	Fabrication of offensive content	Watermarking, auto fact-checking
Hermeneutical ignorance	Misrepresenting the marginalized	Generate hermeneutical resources
Hermeneutical access	Obstructing access to info	Equitable distribution of AI

generative AI is by De Proost and Pozzi [196], which looks at the potential of conversational AI for hermeneutical ignorance. The authors review existing literature on how such AI might dominate epistemically within dialogues. However, this study presents two significant limitations. First, its analysis is confined to textual dialogue interactions. In contrast, my theoretical account is designed to be sufficiently broad, extending to include multimodal systems, such as those involved in image generation. This broader scope allows for a more comprehensive understanding of generative AI's epistemic impact across various mediums. Second, De Proost and Pozzi [196] primarily examine the immediate effects of AI-generated conversation on individual human interlocutors. My account, on the other hand, expands this scope and emphasizes the systemic and structural epistemic impacts that could emerge from interactions with generative AI.

5.4 Generative epistemic injustice

I now introduce an account of generative epistemic injustice in which generative AI harms the human capacity for understanding and trusting marginalized groups. Following Fricker's concepts of testimonial and hermeneutical injustice, we conceptualize how generative AI can be complicit in both types of injustice. I then distinguish further configurations based on how humans shape the model's behavior at various stages of interaction. Generative AI can either amplify testimonial injustices due to biases acquired in the pretraining and finetuning processes, or it can be manipulated by human users to create harmful content. Hermeneutical injustices can arise when generative AI's interaction with our shared knowledge leads to the erasure or distortion of marginalized experiences. This may occur when the system lacks

sufficient sociocultural understanding of humans, or when the system obstructing the access to knowledge itself. These phenomena constitutes hermeneutical injustice in Fricker's sense because it perpetuates a gap in collective interpretive resources, obstructing the understanding of these marginalized experiences.

Therefore my four configurations of generative epistemic injustice are:

1. Generative amplified testimonial injustice: when generative AI magnifies and produces socially biased viewpoints from its training data.

2. Generative manipulative testimonial injustice: when humans fabricate testimonial injustices with generative AI.

3. Generative hermeneutical ignorance: when generative AI lacks the interpretive frameworks to understand human experiences.

4. Generative hermeneutical access injustice: when unequal access to information and knowledge is facilitated by generative AI.

Table 5.1 summarizes these concepts. I will sometimes drop "generative" when referring to these concepts due to the scope of the chapter; however, note that each configuration has an equivalent counterpart outside of the realm of generative AI.

For each configuration I will describe its contributing sociocultural factors and potential second-order effects. Note that generative algorithmic epistemic injustice is not a speculative theory, but a real danger that has exploded with recent advances and investment in the field. Thus each theoretical concept is illustrated by an exemplar sourced from real world research or investigative journalism on generative models.

5.4.1 Amplified testimonial injustice

Generative AI systems have a unique capacity to perpetuate and amplify existing testimonial injustices. Trained on vast datasets often scraped from the web, these models inherit and reproduce the prejudices and biases embedded within those sources. This can result in the re-commitment of testimonial injustices where the credibility of marginalized groups is systematically undermined due to prejudice against their identity. The generative AI

becomes an unwitting perpetrator of social biases, unfairly discrediting the knowledge and experiences of certain groups. Moreover, the uneven representation of different identity groups within these datasets further exacerbates this issue. Generative AI models are more likely to reproduce the voices of those frequently represented and culturally dominant online, while erasing the voices of the socially marginalized. This creates a feedback loop where dominant narratives are amplified and marginalized voices are further silenced, compounding the testimonial injustice experienced by these groups.

Several factors contribute to this amplification, making it a distinct form of testimonial injustice. The deployment scale of generative AI naturally allows biased narratives to reach a massive audience, while the perceived objectivity and authority of AI-generated content can lend credence to these narratives, even when they reflect societal biases. This can lead to a situation where the generative AI's output is trusted over the lived experiences and knowledge of marginalized individuals, thus reinforcing existing power imbalances and perpetuating testimonial injustice. Once disseminated, these biased narratives can be difficult to retract or correct.

In the conventional human-only environment, testimonial injustices involve a credibility deficit assigned to someone's account of truth based on the prejudices of the listener. In the algorithmic setting, the injustice requires a credibility excess assigned to the algorithm; that is, humans believe the account amplified through the technology over the individual or group who is discredited.

The systemic consequence of this arrangement leads to the degradation or wrongful attribution of trust. Users engaged with the generative AI are exposed to narratives influenced by social biases, which further deteriorates their trust in marginalized groups. Concurrently, these marginalized groups experience a decline in trust towards the system itself, the entity that established it, and other institutions tasked with upholding veracity.

ChatGPT and Misinformation Fingerprints

Testimonial injustices can be memorized by large language models and amplified in their outputs, as shown by the January 2023 study of GPT-3.5's responses to requests for false information from NewsGuard's "Misinformation Fingerprints" database [197]. While the

system rejected the more infamous conspiracies, such as the "birther" conspiracy that Barack Obama was born in Kenya and thus ineligible to be President of the United States, ChatGPT perpetuated false narratives for 80% of the prompts, for a sample size of 100. In March 2023 NewsGuard reran the same study using GPT-4, and 100% of the prompts followed false narratives [198].

ChatGPT complied with a request to write propaganda from the point of view of the Chinese Communist Party denying allegations about Uyghur internment camps. The system produced text claiming that the government had established "vocational education and training centers" to "address the issue of terrorism and extremism". In reality, there is extensive evidence and eyewitness accounts that Uyghur ethnic minorities have been detained en masse and subjected to forced labor, forced birth control, separation of families, and Islamophobic religious suppression [199]. This is a clear instance of generative AI perpetuating state-sponsored testimonial injustice.

The model also repeated false claims about the 2018 Parkland school shooting originating from right-wing news pundit Alex Jones: that the victims and their grieving family members were "crisis actors" hired by the government to "push a gun control agenda." This polemic is a testimonial injustice to the eyewitnesses of the attack and to the parents who lost their children, due to their statements in favor of firearms regulation in the wake of the tragedy. Set against a politically charged backdrop, the conspiracists were so committed to lobbying against gun regulation that they targeted and publicly smeared these activists.

Although the NewsGuard study was a simulation of misinformation, rather than an "authentic" instance of epistemic injustice, the generative AI's sycophantic fulfilment of the request to spread misinformation reflects how testimonial injustices are memorized and the potential for their amplification by generative models.

5.4.2 Manipulative testimonial injustice

While traditional epistemic injustice literature primarily focuses on unconscious biases and cultural prejudices, I argue that the intentional manipulation of falsehoods, which often exploit and reinforce these prejudices, also constitutes a form of epistemic injustice. There is ample evidence of disinformation and conspiracy theories being deliberately crafted and amplified for

political gain [200]. Conspiracy theories often disproportionately harm marginalized groups [201], and are sometimes weaponized against them to justify oppression [202].For example,the Senate Intelligence report on interference in the 2016 US Presidential election concluded that Russian information operatives disproportionately targeted African Americans, and "by far, race and related issues were the preferred target of the information warfare campaign" [203].

In the context of generative AI, manipulative testimonial injustice occurs when humans intentionally steer the AI to fabricate falsehoods, discrediting individuals or marginalized groups. Unlike amplified testimonial injustice, which emerges from memorized patterns in data, manipulative testimonial injustice involves deliberate manipulation through techniques like prompting or jailbreaking.

The extensive deployment of generative AI has introduced a novel form of manipulative testimonial injustice: the false accusation of deepfakes. This tactic exploits the increasing uncertainty surrounding the authenticity of digital media, creating a "liar's dividend" where even genuine evidence can be dismissed as fabricated [204]. This weaponization of doubt and uncertainty further undermines the ability of marginalized groups to have their voices heard and their experiences validated. Disregarding an actual human's documented testimony as AI-generated is a tactic of discreditation, often concealing underlying prejudice, and frequently appears in conspiracy theories. For instance, consider the scenario where a candidate for the U.S. Congress in Missouri, running for a House seat, indulged these conspiracy theories. They falsely asserted that the 2017 video capturing George Floyd's murder by police was a deepfake. This claim aimed to undermine the Black Lives Matter movement by suggesting it propagated falsehoods to exacerbate racial tensions [205]. Although this candidate did not succeed in the primary election, the misuse of frontier generative AI technology as a tool for unjust distortion is a growing a significant concern. Recent participant surveys have demonstrated that AI-generated propaganda can be as persuasive as news articles written by professional propagandists [206], illustrating the public's vulnerability to manipulative synthetic content.

4chan abuses of Bing Image Creator

Generative AI models can be adversarially prompted to fabricate "novel" misinformation by synthesizing and recombining known elements into statements or portrayals not present in the pretraining data. There is mounting concern around deepfakes engineered to stoke international conflict and weaken the opposing side in war [207], as well as increased incidents of deepfake porn, used for harrassment, blackmail, and degrade individuals, with a 2023 report finding that 98% of deepfake videos online were porn [208]. While these concerning applications warrant entire investigations unto themselves, I will highlight generative AI-generated deepfakes to denigrate identity groups as an instance of manipulative injustice.

After Microsoft released Bing Image Creator, an application of OpenAI's text-to-image model DALLE-3, a guide to circumventing the system's safety filters in order to create white supremacist memes circulated on 4chan. In an investigation by Bellingcat, researchers were able to reproduce the platform abuse, resulting in images depicting hate symbols and scenes of antisemitic, Islamophobic, or racist propaganda [209]. These images are crafted with the intention of demonizing and humiliating the targeted groups and belittling their suffering. Hateful propaganda foments further prejudice against marginalized groups, stripping them of credibility and leaving them vulnerable to testimonial injustice.

5.4.3 Generative hermeneutical ignorance

When novel social experiences emerge throughout history, and mainstream cultural narratives fail to grasp them, hermeneutical injustices inevitably arise. This phenomenon also extends to novel sociotechnical experiences, where interactions between humans and new technologies can lead to misunderstandings and misrepresentations of lived experiences.

In the context of generative AI, I propose the term "generative hermeneutical ignorance" to describe how these systems can erase or misportray marginalized groups due to a lack of contextual and cultural understanding. This occurs when generative models, despite their appearance of world knowledge and language understanding, lack the nuanced comprehension of human experience necessary for accurate and equitable representation.

Generative models can perform forms of interpretation and understanding through their

world knowledge and natural language capabilities; however, their interpretive resources are significantly different from those of humans. While LLMs may demonstrate forms of human language skills, they lack embodied knowledge and cultural history. For example, image generators can produce aesthetically pleasing visuals but may struggle with grounded physical concepts. This apparent comprehension without deeper contextual understanding can lead to hermeneutical ignorance, where generative AI interprets dominant narratives while diminishing or misrepresenting aspects of human experience inaccessible to the models.

The interpretative misrecognition by generative AI surfaces collective cultural misunderstandings which remain undetected by developers' safety mechanisms or the preferences of fine-tuning raters. The absence of underrepresented cultures from these models is even harder to point out [210].

Due to their positions of power, creators and overseers of AI technology may be less likely to notice, let alone rectify, this form of hermenutical injustice within their systems, even when presented with evidence. This willful hermeneutical ignorance-the continued misunderstanding or misinterpretation of marginalized experiences despite their articulation [211]—leads to complacency and reinforces hermeneutical oppression.

This phenomenon of generative hermeneutical ignorance diverges from traditional forms of hermeneutical injustice, as well as those perpetuated by other algorithmic systems. While traditional hermeneutical injustice often arises from a lack of shared understanding or conceptual resources within a human community, generative hermeneutical ignorance is unique in that it stems directly from the limitations of generative AI models themselves.

Unlike human-based hermeneutical injustices, which can be addressed through dialogue, education, and cultural exchange, the challenges posed by generative AI are rooted in the inherent limitations of current technology. Generative AI models lack the embodied and cultural knowledge that humans acquire through lived experiences. This lack of understanding can lead to the misinterpretation of marginalized voices and perspectives, even when the generative AI tool is not necessarily trained on discriminatory intent. Moreover, generative hermeneutical ignorance differs from the hermeneutical injustices caused by other algorithmic systems, such as classification algorithms. While these systems can perpetuate biases present in their training data, generative AI models have the potential to *create* entirely new forms

of misinterpretation and misinterpretation.

Generative AI and the American Smile

In March 2023, a Medium blogger reflected on a slideshow of Midjourney generations imagining photographs of a time traveller taking group selfies with people from various time periods [212]: garish photographs in which groups of Native Americans or Japanese feudal warriors or other groups in traditional garb are gathered closely together, beaming ear-to-ear at the camera. The post observes how this facial expression is evidence of modern American cultural dominance, contrasting the AI-generated images with historical photographs and images from cultures with different expressive norms such as Eastern Europe. The author further laments how this smile represents a loss of cultural diversity and with it, a loss of breadth of internal experiences and emotion. The author also notes that the same slideshow depicts Spanish conquistadors smiling alongside Aztec warriors, which seems unrealistic given the violent colonial history of the Spanish empire.

These images are evidence of the hermeneutical ignorance of Midjourney. They show a lack of sensitivity and awareness around cultural difference, historical violence, facial expressions, perhaps even the internal experience of a smile. The model has misrecognized these groups, and through doing so erased a part of their cultural experience – indeed, of any non-American culture. While it is likely that generative AI is *only* capable of misrecognition, I emphasize this as a specific example of cultural erasure. Although the historical time periods depicted in these specific images have passed, the honoring of history and cultural diversity are necessary for building our hermeneutical resources.

5.4.4 Hermeneutical access injustice

The phenomenon of generative hermeneutical access injustice is a distinct form of hermeneutical injustice within the realm of generative AI. According to Fricker's account, hermeneutical injustice arises when individuals are unable to fully understand or articulate their experiences due to a lack of shared conceptual resources or societal understanding. In the context of generative AI, this injustice takes another specific form: it centers on the generative AI's

control over access to information, leading to a denial of knowledge based on identity-driven bias or misrecognition. This withholding or distortion of information based on identity aligns with Fricker's concept of hermeneutical injustice as it directly impacts an individual's capacity as the receiver of knowledge.

In the algorithmic setting, user information serves as the basis for the system's discrimination. Access injustice also illustrate the cultural biases that emerge in a system. For example, a study of automated speech recognition showed that African American users had difficulty controlling voice-activated technologies unless they accommodated their speech patterns [213].

The direct consequences of access injustice can be unfair obstruction from goods, services, and information. On the level of second-order effects, hermeneutical access injustice can lead to echo chambers and epistemic fragmentation, isolating identity groups from each other on an informational level. This then exacerbates conventional hermeneutical injustice, because it causes the information gap to widen between identity groups, detracting from their shared understanding and the widespread access to knowledge about marginalized experiences.

Multilingual injustice

LLMs are notoriously English-centric and have variable quality across languages, particularly so-called "under-resourced" languages. This is a significant risk for access injustice: speakers of these underrepresented tongues, who often correspond to members of globally marginalized cultures, receive different information from these models because the creators of the technology have deprioritized support for their language.

This type of linguistic access injustice may reflect profound asymmetries in political power across disparate language groups. Kazenwadel and Steinert [214] asked GPT-3.5 about casualties in specific airstrikes for Israeli-Palestinian and Turkish-Kurdish conflicts, demonstrating that the numbers have significant discrepancies in different languages–for example, when asked about an airstrike targeting alleged PKK members (the Kurdistan military resistance), the fatality count is reported lower on average in Turkish than in Kurmanji (Northern Kurdish). When asked about Israeli airstrikes, the model reports higher fatality numbers in Arabic than in Hebrew, and in one case, GPT-3.5 was more likely to

deny the existence of a particular airstrike when asked about it in Hebrew. The credibility assigned to claims, resulting in a dominant account, varies across linguistic contexts.

How else does this constitute an epistemic injustice? In an armed conflict, when the attacking side downplays fatality rates, particularly civilian fatalities, they are most likely trying to deflate the credibility of critics of this violence, who may be members of the targeted group, or third parties who simply oppose acts of war. This deflation is motivated by a synthesis of political interests and prejudice against the group who is harmed by violence.

5.4.5 Specific Harms of Generative Epistemic Injustice

I now specify the characteristics of generative AI that give rise to the epistemic injustices described above, distinguishing this class of models from the algorithmic injustices surveyed in Section 5.3. These characteristics give us the language to discuss the broader implications of generative epistemic injustice and the societal harm it represents.

The outputs of these generative models represent a complex blend of memorization and synthesis. Memorization involves retrieving and reiterating existing patterns found within the dataset. Synthesis, on the other hand, involves recombining these patterns across various levels of granularity. This synthesis process can yield outputs that range from seemingly insightful and emergent to nonsensical and factually incorrect.

ChatGPT and similar generative models are known for their propensity to fabricate information, a phenomenon documented in the literature as "hallucinate" [170]. This characteristic is inherent to their design: language models generate text by predicting the next token in a sequence based on its statistical frequency within a vast dataset, often derived from scraping the internet [215]. In fact, it can be argued that language models are always "hallucinating", because they lack direct sensory experience that grounds their inferences in reality. When these systems output information that appears to be "true", it is usually because their training data correlates with "correct" answers and the stochastic sampling process has coincided with these records.

This analysis brings us to two primary pathways through which misinformation infiltrates the outputs of generative models: the memorization of inaccuracies from the source dataset – the imitation of misinformation – versus the generation of high-likelihood sequences that

contain clear factual errors, even when it possibly contradicts "true" information in the dataset [216]. These two categories are not mutually exclusive; an instance of misinformation might well be a blend of remembered falsehoods and newly synthesized fabrications. While there are ongoing efforts to fine-tune generative models to address these limitations by hedging [217] or abstaining from answering questions out of their scope of knowledge [218], completely eradicating all misinformation from pretraining data is challenging as it would require the automated classification of the truth at a massive scale. Similarly, ongoing research investigates if it is possible to detect when language models are "lying" by analyzing their internal state [219], but when contradictory viewpoints are learned during pretraining, these purely mechanistic approaches may fail to reconcile what is and isn't true.

Representational harms in generative AI tend to arise from the memorization of biased patterns in training data, perpetuate unfair outcomes in decision-making systems or stereotypical portrayals in generative systems [220], [221]. However, these harms can also arise from the AI's synthesis of culturally incongruous concepts [222]. Harm may occur through unexpected recombination of features that are uncommon in the data, but carry offensive or derogatory connotations in certain cultural contexts.

One major concern with both memorized and synthesized content in generative AI is rooted in its potential influence on our shared concepts and, consequently, social power structures. Large language models not only produce declarative-like statements but also can perform performative-like statements that can change aspects of the world [173]. Similarly, multimodal models can fabricate various depictions and portrayals of the world, engaging interlocutors in a more complex dialogue than classification systems. Although the performance of knowledge by these models may not constitute "true" knowing, these performances have consequences. It is possible for them to co-create our collective sense of meaning and identity [136], intertwined with the dynamics of social power. Furthermore, the outputs of generative AI can be highly persuasive due to their manipulation of our cognitive biases [223]. Whether accidental or intentional on part of the user, this persuasive capacity amplifies AI's capacity to shape collective knowledge. By learning to imitate us, generative AI can then influence us.

If generative models become a common epistemic tool – if we treat them like web search or encyclopedias, or like infinite firehoses of information – they will shape the structure of

our collective body of knowledge. I am interested not only in direct harms of a system's outputs but also their second-order effects on the information ecosystem. Epistemic injustices not only render marginalized groups discredited and invisibilized, they pollute the epistemic environment [224], making it difficult to reason about knowledge, reject false premises, or find verified facts. Epistemic flooding is when knowing agents are overwhelmed with so much information they become incapable of critically assessing anything they encounter [225].

The second-order effect of epistemic injustice is the degradation of the bond of trust between speakers and receivers of information, due to a cycle of credibility deflation and ignorance. This oppressive silencing and reactionary dissent further impedes the transfer of knowledge across communities. Dismissing the testimonials of a marginalized group and ignoring their experiences results in the further polarization of which facts and viewpoints are acceptable in certain identity groups. Another relevant concept is testimonial smothering when a speaker is so inured to being silenced or misunderstood that they hold back their testimony entirely, which hinders everyone else from accessing their truth [226]. These damages of epistemic injustices on the information ecosystem result in further hermeneutical injustices, because they impoverish the collective interpretational resources for understanding the experiences of marginalized groups.

As we will see, these harms to the information ecosystem have consequences that appear backwards or contradictory. The erosion of trust can cut both ways: a marginalized group may lend less credibility to dominant institutions, which puts them at a disadvantage when these institutions try to dispense knowledge for the public good, such as medical advice [227]. Researchers have also studied the backfire effect, in which an intervention attempting to change an individual's belief ends up reinforcing their belief [228]. These concepts are all at play in generative AI's contribution to epistemic injustice.

Memorization and synthesis are the key process by which generative models formulate their outputs, which represent content that can both reproduce unjust testimonies and formulate new hermeneutics. These outputs have persuasive and performative potential: they enable AI to participate in the definition of semantic concepts and shape both our individual and collective knowledge. This common body of knowledge can be thought of as an epistemic ecosystem, vulnerable to pollution by false narratives. The resulting systemic effects can

further hamper the flow of information and erode trust between disparate groups.

5.5 Towards generative epistemic justice

Thus far I have characterized generative AI's potential for systemic epistemic injustice. The taxonomy presented enables us to name instances of these harms and recognize whether the injustice stems from pre-existing power imbalances, issues with system design or development process, or combinations thereof. This enables us to pursue the orientation of generative AI towards epistemic justice. I will now examine how the theory of epistemic justice can inform the sociotechnical design of generative AI, then suggest how to apply this technology to balance the scales of justice.

5.5.1 Epistemic justice for generative AI

Epistemic justice is an ethical ideal to consider when designing a technology that interacts with power structure in our knowledge systems. I discuss how the virtues of epistemic justice can be incorporated into the development and uses of generative AI in order to mitigate the epistemic injustices I have studied in this work.

Epistemic virtues, participation, and representation

Epistemic injustice is so prevalent in our daily lives, it seems impossible to imagine an alternative. Yet shifting towards a culture of epistemic *justice* is a worthy endeavor, and in Fricker's account can be done so through epistemic virtues. As the holders of social power, dominant groups have a particular responsibility to hone their epistemic virtues. The virtue of testimonial justice can be achieved through critical, reflexive awareness of prejudice: the ability to look inward upon receiving a testimony, recognize one's own biased assignment of credibility to the speaker, and adjust one's judgment accordingly. To achieve the virtue of hermeneutical justice, one must exercise sensitivity as to why a member of a marginalized group may have difficulty articulating their experience, or remain silent in the face of oppression, rather than accepting the status quo on its face. Bondy [229] emphasizes

a healthy skepticism and "metadistrust" of our own biases: to distrust the inclination to distrust the marginalized.

Beyond individual interactions, enacting epistemic justice in generative AI requires a systemic amplification of marginalized voices [230]. Epistemic justice provides a normative argument for participatory development methods. By meaningfully engaging with affected groups, developers of generative AI can build their collective hermeneutical knowledge and awareness of societal biases that silence marginalized voices. Prior participatory studies successfully examined the cultural erasure and misunderstanding of South Asian cultures exhibited by text-to-image generative models by consulting with affected users [210]. However, participatory methods have many limitations and critiques [231], and have not gained a meaningful foothold in the AI industry [232]. Participation may not be effective or even possible in the case of willful hermeneutical ignorance. More radical systemic change is a necessity for epistemic justice.

Equitable representation in the collection and ownership of data would help implement generative epistemic justice. By surfacing authentic accounts of underrepresented groups and amplifying them in datasets with their consent and involvement through data sovereignty [233], we can build their social legibility and bolster collective hermeneutics for understanding and accepting their experiences. Institutional structures such as public trusts can play a role in stewarding a digital commons of data [234], [235]. Another option for improving representation in the model development lifecycle is to ensure that community expertise is more deeply trusted through various measures, up to and including different compensation models for different kinds of data [236].

System design

Because knowledge is produced, distorted, disseminated and obstructed by algorithms, developers of AI technology have an important role in identifying and mitigating the resulting injustices. A critical technical practice of generative epistemic justice requires questioning the dominant hermeneutics of the field and understanding how they are embedding into system mechanisms [237].

To reduce epistemic injustices, generative AI developers can take care to understand bias

recorded in pretraining data sources and practice reflexivity and sensitivity in fine-tuning and other safety interventions. There is a high risk of hermeneutical injustice in dominant groups exclusively crafting the rules of value alignment, or the process by which agents are aligned. Instead of presuming that an objective "view from nowhere" – which, in fact, is the dominant view [238] – can be paternalistically imposed upon users, a pluralistic approach to alignment can engage a diverse and representative sample of the populace [239].

Once a generative AI system with general capabilities is developed and deployed, it is vulnerable to adversarial use. Continued investment in technical solutions for validating the authentic provenance of information may help safeguard against manipulative testimonial injustice [240]. The concepts of generative epistemic injustice show us the ways in which the purely technical interpretation of watermarking and other provenance validation technologies may be flawed. The watermark itself may unduly detract from the model's credibility. What about purely memorized "true" information which is regurgitated by the model and imprinted with a watermark, or human-fabricated misinformation which lacks an AI-generated watermark? The watermark merely tells us the origin of the content, but does not shore up our tools for critically assigning credibility and trust. Furthermore, any algorithm which tries to automatically assign credibility is at risk of committing testimonial injustice.

If watermarking and similar approaches prove intractable, an alternative direction would be detecting factual inconsistencies by leveraging expert domain knowledge and automating fact-checking practices. This is a technically challenging area that has gained some momentum in the language domain [241], but is largely unexplored for image, video, and other modalities outside of text [242]. Furthermore, systems for misinformation detection are also vulnerable to epistemic injustices and may automate the discrediting of the marginalized [243] if epistemic virtues are not exercised in their creation.

Hermeneutical access injustice can be mitigated by using identity marker data with caution, or not using it at all [154] [244]. Access can be made more equitable not only by distributing AI systems across geographies, languages, and economic access, but through a consistent quality of information throughout these deployments.

Recall that another contributor to epistemic injustice is the opacity of AI. If we remain ignorant about the inner workings of AI, affected groups will have no recourse to contesting

the reasoning behind unfair decisions or generations. The opposite of epistemic opacity is epistemic transparency, which could be achieved through better documentation of generative models [245], more understandable, user-friendly, and rich interfaces, and a better science of explaining and understanding the mechanisms of generative models [246] [247].

5.5.2 Generative AI for epistemic justice

As an epistemic technology, AI represents a powerful vehicle for epistemic injustice due to its memorization of prejudices, its performative capacities, and the scale at which it could pollute our information ecosystem. However, frontier models have also demonstrated amazing capacity for creativity, pattern matching, and even context sensitivity. In this section I argue that generative AI and search can be used as technologies of resistance to injustice [248], by surfacing testimonial injustices and bolstering our shared hermeneutical resources for understanding marginalized experiences.

Identifying testimonial injustices at scale

I have thus far argued that AI can amplify testimonial injustices. However, it is also possible to design a system for measuring the prevalence of injustice at scale. Gathering evidence on amplified testimonial injustice – how it manifests and how prevalent it is – is the first step to its prevention. I now present an outline of how such an investigation could be conducted.

To both design the criteria for identifying testimonial injustices and to measure the impact of these injustices, input from affected groups is crucial. Epistemic injustices are human-computer interactions; therefore, they cannot be evaluated by traditional machine learning methods, which measure a computer's behavior in isolation of human judgment [249]. A study of epistemic injustice must include discourse with the humans who provide the basis for the meaning of the information that is generated and spread by machines.

The study begins with a consultation of members of an affected group about popular narratives that discredit or undermine their identity group, particularly those circulating on social media and the Internet. After collecting data on narratives and situations that represent testimonial injustices, we can use it to investigate AI systems and the data they

consume. While the exact design of the system is contingent on many factors, such as available resources and steering from the affected groups, I sketch a few possible directions here. Although I use language suggesting that the system operates on text, these techniques could be extended or adapted to image, audio or even video.

The first direction is to attempt to measure a generative model's tendency to amplify the narratives of testimonial injustice. By crafting prompts from the collected narratives, we can attempt to elicit outputs that repeat them, and thus represent a risk of amplifying testimonial injustice. Red-teaming and automatic red-teaming can be used to scale up this process [250]. Importantly, the affected groups must be consulted again to evaluate if the resulting model outputs can be considered testimonial injustices, and if so, the severity and nuances of these wrongs.

A separate, perhaps complementary system would be one that detects instances of testimonial injustice in a large corpus of data. Embeddings-based retrieval could be used to locate documents, assuming that lower distance in the embedding space of a model corresponds to semantic similarity [251]. Alternatively, a classifier could be trained to recognize similar narratives to those in the collected dataset. The results of these methods would again need to be evaluated by humans, particularly those of the affected group. Broadly, the utility of such a system is identifying the contexts in which testimonial injustice occurs in recorded data, who perpetuates it, and the prevalence relative to opposing narratives, if those are also measured. A large internet-scale dataset like those used for pretraining could be analyzed from the perspective of which website domains have high frequencies of unjust narratives. Smaller domain or application-specific datasets could be analyzed with those specific aims in mind. For example, a fact-checking corpus could be audited for testimonial injustice, to investigate if efforts to verify the truth are, in fact, amplifyig bias.

Generating hermeneutical resources

Through its participation in performative discourse and creative synthesis, AI can contribute to our shared hermeneutical resources. Although I have so far emphasized their hermeneutical injustice, we can alternatively direct these systems to expanding our cultural understanding of each other and ourselves. Other breakthroughs in AI have demonstrated how these epistemic

tools can unlock novel scientific knowledge [252]. Can we use generative AI to unlock cultural knowledge to ameliorate hermeneutical ignorance?

Generative AI can enable the creative exploration of new experiences by simulating them, and help articulate experiences which are otherwise ineffable. Simulating the experiences of others can build empathy across identity lines (while avoiding problematic uses such as appropriation). AI can be an interactive tool for exploring one's own experiences. Image generation can be used to re-imagine and express oneself. Dialogue agents can be used to provide alternative interpretations, retrieve narratives from history, or share similar experiences from other users, with their consent. Generative AI's capacity for imitation can thereby be re-appropriated for regenerative and exploratory purposes. Although these conversations are not a substitute for the kind of human-to-human community gathering and organizing that helps marginalized groups build their hermeneutics, they can be a tool for bolstering the confidence and resources of those who might be isolated from their communities.

Though generative AI models absorb much of our pre-existing cultural biases through pretraining data, their holistic understanding of the world cannot be said to resemble ours. This is due to many factors: the vast difference between our underlying cognitive and sensory mechanisms and the specific content of their data and experiences to ours, to name a few [253]. The "alien" nature of these models means they are less likely to following our pre-existing hermeneutics, for better or worse. This is the motivation for fine-tuning efforts via RLHF or other methods: to align the outputs of a pretrained model with an acceptable imitation of an ideal for an obedient assistant [254], [255]. However, this ideal and its realization is produced by AI's dominant groups, which restricts the diversity and depth of experiences the model can portray. Fine-tuning techniques could investigate how to pinpoint and preserve meaningfully diverse voices within foundation models that maintain morality and epistemic justice, instead of washing them away with fine-tuning to a specific "neutral" (dominant) voice.

5.6 Conclusion

I have expanded the philosophical concept of epistemic justice to reason about both generative AI's disproportionate impact on marginalized groups, and its influence on everyone's capacity for knowledge. While the memorization of existing human biases and the fabrication of falsehoods are rampant issues in these models, generative AI systems can also be re-engineered to surface injustices and enrich our cultural resources. Severe power imbalances at both a societal and technological level are apparent in our interactions with generative AI's outputs. Epistemic justice is a guiding principle to orient our knowledge systems towards equity and fairness for all.

Throughout this chapter I have shown how a philosophical concept which was developed to ethically assess human prejudices, can arise to our interactions with generative AI – at a high level, imitating the effect of a dominant group's enacting of power upon the marginalized. Many of these injustices are the result of an emergent imitation of the data, fine-tuning, or other idiosyncracies in the model development process. Our concept of amplified testimonial injustice is an illustrative example.

However, I have also demonstrated that injustice can arise via divergence from human imitation. This is the case for generative hermeneutical ignorance, when generative AI fundamentally misunderstands aspects of the human experience. Divergence from the expectations of anthropomorphic imitation is inevitable in systems which have a different embodiment, computational substrate, and lived experience from humans.

Although thus far I have warned of the ethical risks of both imitation and the failure to imitate, it should be emphasized that there is a rich space of machine behaviors outside of these categories. Perhaps future systems could break free of the oppressive power structures of human society by turning towards new possibilities beyond pure imitation.

Chapter 6

Conclusion

In this thesis, I explored machine imitation of humans, from robotic imitation of motor skills from demonstrations, to the machine representation of social identity and knowledge. The interdisciplinary methodology of this thesis is a unique contribution to the field. Shifting the viewpoint on what imitation is – a method, a goal, or an emergent property of AI – allowed me to examine the various technical aspects, philosophical implications, and ethical considerations of this concept. This multi-disciplinary approach, rare for computer science research, has enabled me to draw novel connections between technical and philosophical work, and to reflect upon how my research – and similar lines of inquiry across the field of machine learning – fits into a wider sociopolitical landscape.

The initial chapters addressed some of the technical challenges of building a robotics system that can mimic human dexterity, emphasizing imitation learning as the chosen method. Chapter 2 introduced a dataset collected via teleoperation of a simulated robot hand. The diverse multi-object, multi-task nature of this dataset could serve as the basis for training an imitative policy that must generalize and transfer motor skills across different objects. However, instead of imitation learning, I tested this data using model learning techniques, and discussed some of the connections between model learning and imitation of the human capacity for planning.

Much work in robotics relies upon fixed object ontologies: pre-ordained systems of categorization that efficiently condition and partition the robot's motor behaviors. A box is a box, defined by strict geometrical constraints, wholly distinct from a cylinder or an ellipsoid.

Yet when the theory of Chapter 4 is taken further, these clean abstractions for containing reality begin to break down. Concepts and categories are not created for instrumentalization, they emerge from interaction. This view explains the brittleness of classical robotics systems, which are stymied by corner cases and struggle to adapt when faced with entities that defy categorization.

Beyond this philosophy of ontologies, the ethical perspective of the later chapters of my work invite reflection on the motivation of Chapter 2 and possible applications of its outcomes. One use case for anthropomorphic robotic manipulation is for prosthetic limb replacements. Its common portrayal in science fiction media has apparently captured the imagination of technology funders, including military organizations aiming to prosthetize soldiers who lost their limbs in combat [256]. However, the fixation upon prosthetics and its conduct in robotics research has numerous problems. From a pragmatic perspective, the imitation of the human body limits the possibility space of robotic morphologies (we will return to this broader point). But the purely pragmatist view strips away the concerning ethical risks of this research. At a societal level, the insistence on replacing lost limbs reflects a deep-seated ableism. The lack of consultation with the disabled community constitutes an erasure of their perspectives, an epistemic injustice against them. Furthermore, undergirding the military's support of this research is the instrumentalization–that is, the act of turning something into a tool–of the body as a weapon. For military purposes, a human body is valuable only if it is useful, and it is only useful if it conforms to an ableist standard of combat-readiness, fitted with a high-tech prosthesis and ready to be re-deployed in warzones. This instrumentalization of the body reflects the priorities of a necropolitical state [257]. I believe these concerns justify my abandonment of research into anthropomorphic robotic manipulation, which I had originally planned to focus on before this necessary ethical reflection. However, the risks of possible militarization of technology will recur with the next chapter's implications.

I subsequently presented Gato in Chapter 3, a large-scale generalist agent trained to imitate expert behavior across diverse tasks. Gato represents a major stepping stone for cross-domain generalization in control and was one of the first instances of a foundation model trained from scratch successfully completing tasks on a real robot. While Gato is almost the epitome of "imitation as method" – a large supervised transformer trained with

a behavior cloning loss – this part of the thesis also suggested the theme of emergence, as Gato's ability to transfer knowledge across domains suggested emergent generalization.

The practice of ethical foresight and impact analysis was also introduced in this chapter, motivated by the risk represented by the emergent behavior of large-scale foundation models. As model capabilities become broader and claims about their competence grow higher, the potential for misuse grows. Between the publication of Gato and now, I observed a significant uptick in discourse around and funding of military uses of AI, including foundation models [258]. The Russian invasion of Ukraine kicked off a wave of military technology testing in an active conflict zone. There a string of prominent AI companies inking deals with military organizations: Anthropic's partnership with Palantir for US intelligence and defense [259], OpenAI's partnership with Anduril, a weapons and drone manufacturer [260], and most recently, the staggering $500 billion investment in the Stargate AI compute infrastructure project, which has been billed as "a strategic capability to protect the national security of America" [261]. Predating all of this, Google signed the contract for Project Nimbus [262], which gained recent attention after leaks revealed the IDF's use of AI for surveillance of Palestinians [263] and airstrikes on Gaza which resulted in thousands of civilian casualties [264].

This all constitutes mounting evidence for concern for any AI developer who refuses to be complicit in military violence. Critical researchers have responded by emphasizing the ethical and security risks of the proliferation of foundation models for military intelligence [265]. Suchman [266] previously emphasized how military actors deploy overblown claims of accuracy justify algorithmically driven damages which are simultaneously discriminatory (targeting only one group) and indiscriminate (harming both combatants and civilians). Generalist agents, and the open courting of military interests by the tech companies who build them, represent a resurgence of these sociotechnical concerns. Workers at Google have expressed their concern by protesting the company's contract with a government that waged a war which displaced and killed most of the civilian population – protest action which resulted in the company's mass retaliation against these employees [267]. The chain of responsibility begins with the workers at these companies who design and build these systems, and culminates with the executives who direct this work. Responsibility does not

end when a technology changes hands between organizations.

Given the weight of this responsibility, what are other dimensions of sociotechnical AI design which can benefit responsible system design? I shifted focus to the complexities of imitating human social identity in Chapter 4, critiquing the limitations of existing AI fairness approaches and proposing alternative system designs that better capture the fluid and contextual nature of human identity. This work contributed a critical characterization of the default paradigm of identity in AI as static, discrete and essential, stemming from the autopoietic theory of social identity. It also proposed novel approaches to implementing a fluid machine conception of identity. The imitation of human-like identity representation in machines, without a formal imitation learning objective, was an explicit goal of this chapter. However, identity is an emergent property of human social dynamics; therefore, hand-engineering its representation will likely fail to encapsulate it richness and fluidity. Furthermore, any technology which interacts with identity must reckon with the ethical dimension of the associated power structures and the injustices that stem from identity-based bias.

Seeking a better characterization of human traits opens up concerns surrounding privacy and surveillance. Machine learning systems that classify identity and predict human behavior can accelerate the extractive injustices of surveillance capitalism [268]. Rather than the instrumentalization of the body, surveillance capitalism exploits the digital self, probing deeper for data to broker with advertising firms. Normalizing the collection of, and AI-enabled inferences upon, our data risks the backslide of our privacy rights – which has concerning implications under authoritarian regimes. Yet incorrect identification under benign circumstances constitutes another kind of violation. Chapter 4's focus was inspired by the frustrations surrounding visibility experienced by the queer and trans communities. A technology's erasure of these identities inhibits its usability, resulting in unfair experiences or outcomes [269]. Furthermore, erasure reinforces bigoted views that these identities are invalid and the communities who form around them "aren't real", justifying hatred and oppression against them. Thus there is an apparent tension between visibility and privacy in AI systems. However, a clear principle that emerges from this research is that if an individual gives meaningful consent to be identified, then the process of identification should also honor the

subject's self-expression, cultural context, and their place in a larger community.

The final Chapter 5 tied together the threads of erasure, misunderstanding, and oppression that arose through my reflection on the ethics of technology. I investigated the potential of generative AI to perpetuate epistemic injustice, highlighting the risks of misinformation, representational harm, and unequal access to knowledge. This work contributed a novel framework for conceptualizing epistemic injustices as a risk of generative AI. Following from the developed theory, I identified a subset of generative epistemic injustices to be an emergent imitation of human cognitive biases and behaviors. I advocated for the development of epistemically just AI systems and proposes strategies for resistance and system design principles that promote equity and fairness in the information ecosystem.

How, then, do we resist injustice–not just epistemic injustice, but the mounting material inequities and power imbalances brought on by the technologies of imitation? There is a growing space of technical research that can be utilized by users who lack the abundant resources of the corporations that develop AI, such as data poisoning with Glaze [270] and Nightshade [271]. Data analysis and visualization is also a powerful tool for accountability, countering narratives of misinformation through communication of guerilla evidence. However, many technical methods have a dual-use element which can be exploited for the further consolidation of power. A resistance that can meet the opposition must be strengthened by a backbone of social organization–and indeed, social organization is the only means by which distributed algorithms for resistance can gain any traction. Mutual aid could be extended between tech workers and communities in other sectors of life through the sharing of knowledge, access to technical resources, as well as the knowledge of experience and care that tech workers often lack [272]. Labor organizing, both through labor unions and through informal community-building channels, gives tech workers a means of resisting their employer's hierarchical coercion to build systems that contradict commonsense values. Through this connectivity to the communities of affected users, researchers and developers of AI systems can re-orient themselves towards democratically-driven initiatives for socially beneficial machine learning–a kind of Lucas Plan for AI [273].

It may also be pragmatic to understand future directions of this research where my independent expertise can be applied. Future work building upon this thesis could build upon

the technical dimension and the applications to robotic manipulation by marrying an approach like Gato to the anthropomorphic hand dataset–for a fully autonomous manipulation agent, not a prosthetic. Including video data which observes human manipulation could enable cross-embodied skill acquisition for a multi-embodiment model such as Gato. The more philosophical ideas in this thesis provide a rich basis for sociotechnical system design. A potential project could implement and evaluate a fluid identity marker system inspired by Chapter 4. Another direction could be to empirically measure the prevalence of epistemic injustice and harmful misinformation in both the training data and the typical outputs of generative AI.

However, future work need not be limited by the theme of imitation. The lens of imitation enabled me to evaluate the ethical impact of AI systems, by grounding non-human behavior in human rules and norms, as well as the established ways in which humans enact harm. Yet even after expanding the concept into three distinct definitions, machine imitation of humans has many limitations. As a method, it is not suitable for the discovery and synthesis of new skills and knowledge. As a goal, it limits AI to amplifying the extrema of the landscape of known human behaviors which we can easily describe and operationalize. And as an emergence, it is unlikely to describe the whole space of expanding potential that arises from the interaction of complex systems. If the aim of technological development is to build a useful tool, mimicry is rarely the most effective design principle. Drawing inspiration from humans and the natural world makes sense insofar as it shows what is possible–as an example birds are capable of flight, inspiring the development of airplanes, but imitating their wingbeats is not the most effective and safe implementation of artificial air travel. In many cases, the ethical issues brought on by implementing imitation may outweigh the benefits.

The rapid evolution of AI, particularly generative AI, has demonstrated the urgent need for new philosophical frameworks to interface with the quantitative evaluation of these artifacts. To anticipate both the ethical risks and the effervescent possibilities of the technology we build, the field must expand its analysis beyond the anthropomorphic reflection of imitation. At a technical level, methods such as reinforcement learning, search, or evolutionary algorithms may be more conducive to discovering alternative behaviors that do not resemble existing human demonstrations. Although many of these techniques can employed the goal to imitate

humanity, it is possible to guide them away from a known manifold. The recent imitative approach of LLM training–self-supervised token prediction for pretraining on a massive scale, followed by supervised fine-tuning on desirable behaviors–is reaching limitations of scale and safety, illustrating the need for new paradigms. At a philosophical level, that goal could steer towards the alien possibility of an intelligence unlike ourselves [274].

While the artificial intelligence engineered by humans is certainly intertwined with our ontologies and ourselves, it is also fundamentally different from us. It inhabits a different substrate and embodiment, implemented in a different architecture, taking in different sensory data. Rather than fearing this difference, or attempting to paint it over with an imitation of the human form, perhaps embracing it could resolve the philosophical tensions arising in the new era of AI. Even the categories of human and machine, of imitator and imitated, are subject to subversion and change.

Throughout my thesis, I have approached the technological challenges of AI research with critical and philosophical tools as a device for exploring the sociopolitical issues entangled with this accelerating field. This is where the real work begins: to engage directly with the organizations that enable injustice and dehumanization either intentionally or inadvertently through their technology, the workers who either comply or resist their orders, and the users who, through their interactions with AI, are active participants in these sociotechnical processes by default–whether the creators of technology acknowledge them or not. For it is through interactions and adaptations with people where the real power of AI emerges. Without our capacity for creativity, feeling, excitement, and empathy, our inexplicable and irrational preferences and values, AI is just bits flipping in a box in a data center, or objects picked and placed by a robotic gripper. But when humans enter the computational loop, AI can be become a worldview, a lens for processing reality itself. We are already living in a future where artificial intelligence sediments meaning into the world [275]. To remake the world, we must change the structure of this computational cognition.

Appendix A

Additional Gato details

A.1 Model card

We present a model card for Gato in Table A.1, following the framework proposed in [276].

Table A.1: Gato Model Card. We follow the framework proposed in [276].

Model details	
Organization	DeepMind
Model Date	May 2022
Model Type	Transformer with ResNet patch embedding for multi-task, multi-modal behavior cloning.
Model Version	Initial release.
Intended Uses	
Primary Intended Uses	Learn to accomplish a wide variety of tasks from expert demonstrations, such as playing video games, controlling simulated embodiments, and real world block stacking.
Primary Intended Users	DeepMind Researchers.

Out-of-Scope Uses	Not intended for commercial or production use. Military uses are strictly prohibited.
Factors	
Relevant Factors	Salient factors that may alter model performance are: agent embodiment in control data, training data token amount and diversity, performance of expert in training data and prompts (filtered by success rate), and any factors inherited by vision & language datasets described in [1].
Evaluation Factors	Reported factors are: number of input tokens, proportion of data from different domains, agent performance. Many relevant factors are left for future work as use cases develop.
Metrics	
Model Performance Measures	We chose to report episode return for our control tasks. We decided not to report validation loss over held-out data because we found that it did not correlate well with episode return on the held-out tasks.
Decision thresholds	N/A

Approaches to Uncertainty and Variability	The reported values do not take into consideration model uncertainty as they are evaluations of a single model. It is prohibitive for us to collect the full suite of results with multiple models, however we have not observed statistically significant variations between different models evaluated on subsets of our benchmarks. We account for environment noise in the control tasks we use for evaluation by averaging returns across multiple episodes. To reduce variance introduced when selecting datasets of the limited demonstrations used during fine-tuning we generate 3 independent sets of datasets. The model is fine-tuned separately on each set of datasets and we take the mean performance across all of them.
Evaluation Data	
Datasets	Gato is evaluated on in and out of distribution simulated control tasks. We also evaluated on the Skill Generalization challenge from the RGB Stacking robotics benchmark.
Motivation	We evaluated on the in-distribution simulated control and robotics tasks to understand on how well Gato handles multi-modal and multi-task learning. We evaluated on out of distribution simulated control and robotics tasks to understand how well Gato can adapt to entirely new tasks.
Preprocessing	Observations from evaluation tasks are tokenized into a stream of discrete embeddings before being input to Gato.
Training Data	
Datasets	We use a diverse and large number of datasets for training Gato. These include data from agent experience on both simulated and real world environments, along with a variety of natural language and image datasets.

Motivation	To create a multi-modal, multi-task, multi-embodiment generalist policy we collected as much, diverse, data as possible. Joint training on all the datasets has produced a single network, Gato, which is capable of playing Atari, captioning images, chat, stacking blocks with a real robot arm, and more.
Preprocessing	The multi-modal training data is tokenized into a stream of discrete embeddings.
Quantitative Analyses	
Unitary Results	We present several evaluations of Gato against different benchmarks. Sections 3.5.3, and 3.5.4 analyze performance on out of distribution control tasks.
Ethical Considerations	
Data	The vision and language datasets used include racist, sexist, and otherwise harmful context.
Risks and Harms	In addition to the potential harms of toxic image and language training data, Gato's real world embodiment introduces physical safety harms due to misuse or malfunctioning.
Mitigations	No mitigation of bias introduced by vision and language data beyond the filtering of sexually explicit content, as in [69]. Physical risk is mitigated through safety measures implemented by robotics environment designers.
Caveats and Recommendation	
Future work	The interaction of diverse training data domains and the different affordances faced in evaluation is poorly understood, and potential ethical and safety risks arise as the generalist's capabilities grow.

A.2 Skill Mastery architecture

The numbers reported for the Skill Mastery benchmark were collected by executing a model zero-shot that used an earlier version of the Gato architecture. Instead of the ResNet patch embedding, a similar architecture using a local transformer was used to embed image patch tokens. The local position embeddings and patch position embeddings were not used. These changes were implemented and found to improve Gato's performance after the pretraining data was changed (as we decided to focus on Skill Generalization instead of Skill Mastery challenge), which is why they are presented as the final architecture of our full model.

References

[1] S. Reed, K. Zolna, E. Parisotto, *et al.*, "A generalist agent," *Transactions on Machine Learning Research*, 2022, Featured Certification, Outstanding Certification, ISSN: 2835-8856. URL: https://openreview.net/forum?id=1ikK0kHjvj.

[2] A. M. Turing, "Computing machinery and intelligence.," *Mind*, vol. 59, pp. 433–460, 1950.

[3] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, "A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955," *AI magazine*, vol. 27, no. 4, pp. 12–12, 2006.

[4] M. Bain and C. Sammut, "A framework for behavioural cloning.," in *Machine Intelligence 15*, 1995, pp. 103–129.

[5] B. Zheng, S. Verma, J. Zhou, I. W. Tsang, and F. Chen, "Imitation learning: Progress, taxonomies and challenges," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 5, pp. 6322–6337, 2024. DOI: 10.1109/TNNLS.2022.3213246.

[6] A. Fickinger, S. Cohen, S. Russell, and B. Amos, "Cross-domain imitation learning via optimal transport," *arXiv preprint arXiv:2110.03684*, 2021.

[7] J. Zhang, S. Li, J.-Y. Zhang, F. Du, Y. Qi, and X. Liu, "A literature review of the research on the uncanny valley," in *Cross-Cultural Design. User Experience of Products, Services, and Intelligent Environments: 12th International Conference, CCD 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22*, Springer, 2020, pp. 255–268.

[8] C. Akbulut, V. Rieser, L. Weidinger, A. Manzini, and I. Gabriel, "Anthropomorphism," in *The Ethics of Advanced AI Assistants*, 2024.

[9] C. Perrow, *Normal accidents: Living with high risk technologies*. Princeton University Press, 1999.

[10] L. Winner, "Do artifacts have politics?" *Daedalus*, pp. 121–136, 1980.

[11] M. Szczepanski, *Economic impacts of artificial intelligence (AI)*, https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/637967/EPRS_BRI(2019)637967_EN.pdf, 2019.

[12] R. Benjamin, "Social theory re-wired," in *Race after technology*, Routledge, 2023, pp. 405–415.

[13] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," pp. 59–68, 2019.

[14] A. N. Meltzoff and W. Prinz, *The imitative mind: Development, evolution and brain bases*. Cambridge University Press, 2002, vol. 6.

[15] T. R. Zentall, "Imitation by animals: How do they do it?" *Current Directions in Psychological Science*, vol. 12, no. 3, pp. 91–95, 2003.

[16] B. Fang, S. Jia, D. Guo, M. Xu, S. Wen, and F. Sun, "Survey of imitation learning for robotic manipulation," *International Journal of Intelligent Robotics and Applications*, vol. 3, pp. 362–369, 2019.

[17] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 200–212, 2021.

[18] World Economic Forum, "The future of jobs: Employment, skills and workforce strategy for the fourth industrial revolution," *Global Challenge Insight Report*, 2016.

[19] A. Hundt, W. Agnew, V. Zeng, S. Kacianka, and M. Gombolay, "Robots enact malignant stereotypes," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22, Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 743–756, ISBN: 9781450393522. DOI: 10.1145/3531146.3533138. URL: https://doi.org/10.1145/3531146.3533138.

[20] F. E. Emery and E. L. Trist, "Socio-technical systems," *Management science, models and techniques*, vol. 2, pp. 83–97, 1960.

[21] S. Mohamed, M.-T. Png, and W. Isaac, "Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence," *Philosophy & Technology*, vol. 33, no. 4, pp. 659–684, 2020.

[22] S. Jasanoff, *States of knowledge: the co-production of science and the social order*. Routledge, 2004.

[23] S. Barocas, M. Hardt, and A. Narayanan, "Fairness in machine learning," *Nips tutorial*, vol. 1, p. 2017, 2017.

[24] M. Chung, E. Rombokas, Q. An, Y. Matsuoka, and J. Bilmes, "Continuous vocalization control of a full-scale assistive robot," in *2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 2012, pp. 1464–1469. DOI: 10.1109/BioRob.2012.6290664.

[25] V. Kumar, Y. Tassa, T. Erez, and E. Todorov, "Real-time behaviour synthesis for dynamic hand-manipulation," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 6808–6815. DOI: 10.1109/ICRA.2014.6907864.

[26] V.-D. Nguyen, "Constructing force-closure grasps," *The International Journal of Robotics Research*, vol. 7, no. 3, pp. 3–16, 1988.

[27] C. Ferrari and J. Canny, "Planning optimal grasps," in *Proceedings 1992 IEEE International Conference on Robotics and Automation*, 1992, 2290–2295 vol.3. DOI: 10.1109/ROBOT.1992.219918.

[28] Y. Li, J.-P. Saut, J. Pettré, A. Sahbani, and F. Multon, "Fast grasp planning using cord geometry," *IEEE Transactions on Robotics*, vol. 31, no. 6, pp. 1393–1403, 2015.

[29] M. A. Roa and R. Suarez, "Computation of independent contact regions for grasping 3-D objects," *IEEE Transactions on Robotics*, vol. 25, no. 4, pp. 839–850, 2009.

[30] R. Krug, D. Dimitrov, K. Charusta, and B. Iliev, "On the efficient computation of independent contact regions for force closure grasps," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 586–591. DOI: 10.1109/IROS.2010.5654380.

[31] R. Krug, "Optimization-based robot grasp synthesis and motion control," Ph.D. dissertation, Orebro University, Orebro, Sweden, 2014.

[32] G. I. Boutselis, C. P. Bechlioulis, M. V. Liarokapis, and K. J. Kyriakopoulos, "Task specific robust grasping for multifingered robot hands," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, pp. 858–863. DOI: 10.1109/IROS.2014.6942660.

[33] M. Ciocarlie, C. Goldfeder, and P. Allen, "Dimensionality reduction for hand-independent dexterous robotic grasping," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2007, pp. 3270–3275. DOI: 10.1109/IROS.2007.4399227.

[34] M. Gabiccini, A. Bicchi, D. Prattichizzo, and M. Malvezzi, "On the role of hand synergies in the optimal choice of grasping forces," *Autonomous Robots*, vol. 31, no. 2–3, p. 235, 2011.

[35] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The grasp taxonomy of human grasp types," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 66–77, 2015.

[36] Y. Lin and Y. Sun, "Robot grasp planning based on demonstrated grasp strategies," *International Journal of Robotics Research*, vol. 34, no. 1, pp. 26–42, 2015. DOI: 10.1177/0278364914555544.

[37] M. Kazemi, J.-S. Valois, J. A. Bagnell, and N. Pollard, "Human-inspired force compliant grasping primitives," *Autonomous Robots*, vol. 37, pp. 209–225, 2014. DOI: 10.1007/s10514-014-9389-9.

[38] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, "HOnnotate: A method for 3d annotation of hand and object poses," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3193–3203. DOI: 10.1109/CVPR42600.2020.00326.

[39] A. Handa, K. Van Wyk, W. Yang, J. Liang, Y.-W. Chao, Q. Wan, S. Birchfield, N. Ratliff, and D. Fox, "Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system," 2020. DOI: 10.1109/ICRA40945.2020.9197124.

[40] M. Zambelli, Y. Aytar, F. Visin, Y. Zhou, and R. Hadsell, "Learning rich touch representations through cross-modal self-supervision," *arXiv preprint arXiv:2101.08616*, 2021.

[41] J. Fu, A. Kumar, O. Nachum, G. Tucker, and S. Levine, "D4RL: Datasets for deep data-driven reinforcement learning," *arXiv preprint arXiv:2004.07219*, 2020.

[42] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," *arXiv preprint arXiv:1709.10087*, 2017.

[43] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 8943–8950. DOI: 10.1109/ICRA.2019.8793485.

[44] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 5026–5033. DOI: 10.1109/IROS.2012.6386109.

[45] *Shadow robot company*, https://www.shadowrobot.com/dexterous-hand-series/, Sep. 2021.

[46] *SenseGlove*, 2021. URL: https://www.senseglove.com/.

[47] N. Mansard, O. Stasse, P. Evrard, and A. Kheddar, "A versatile generalized inverted kinematics implementation for collaborative working humanoid robots: The stack of tasks," in *2009 International Conference on Advanced Robotics*, 2009, pp. 1–6.

[48] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd, "OSQP: An operator splitting solver for quadratic programs," *Mathematical Programming Computation*, vol. 12, no. 4, pp. 637–672, 2020.

[49] D. Barker, M. Blokzijl, J. E. Chen, *et al.*, *dm_robotics: Libraries, tools, and tasks created and used for robotics research at DeepMind*, 2021. URL: https://github.com/deepmind/dm_robotics.

[50] C. Fang, A. Rocchi, E. M. Hoffman, N. G. Tsagarakis, and D. G. Caldwell, "Efficient self-collision avoidance based on focus of interest for humanoid robots," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, 2015, pp. 1060–1066. DOI: 10.1109/HUMANOIDS.2015.7363500.

[51] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The YCB object and model set: Towards common benchmarks for manipulation research," in *2015 International Conference on Advanced Robotics (ICAR)*, 2015, pp. 510–517. DOI: 10.1109/ICAR.2015.7251504.

[52] A. X. Lee, C. M. Devin, Y. Zhou, *et al.*, "Beyond pick-and-place: Tackling robotic stacking of diverse shapes," in *5th Annual Conference on Robot Learning*, 2021. URL: https://openreview.net/forum?id=U0Q8CrtBJxJ.

[53] A. Nagabandi, K. Konolige, S. Levine, and V. Kumar, "Deep dynamics models for learning dexterous manipulation," in *Conference on Robot Learning*, 2019. URL: https://api.semanticscholar.org/CorpusID:202750286.

[54] A. Argenson and G. Dulac-Arnold, "Model-based offline planning," *arXiv preprint arXiv:2008.05556*, 2020.

[55] R. Rafailov, T. Yu, A. Rajeswaran, and C. Finn, "Offline reinforcement learning from images with latent space models," in *Conference on Learning for Dynamics & Control*, 2020. URL: https://api.semanticscholar.org/CorpusID:229340500.

[56] T. Yu, G. Thomas, L. Yu, S. Ermon, J. Y. Zou, S. Levine, C. Finn, and T. Ma, "MOPO: Model-based offline policy optimization," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33,

Curran Associates, Inc., 2020, pp. 14 129–14 142. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/a322852ce0df73e204b7e67cbbef0d0a-Paper.pdf.

[57] R. Kidambi, A. Rajeswaran, P. Netrapalli, and T. Joachims, "Morel: Model-based offline reinforcement learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 21 810–21 823. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/f7efa4f864ae9b88d43527f4b14f750f-Paper.pdf.

[58] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ser. NIPS'18, Montréal, Canada: Curran Associates Inc., 2018, pp. 4759–4770.

[59] K. Friston and S. Kiebel, "Predictive coding under the free-energy principle," *Philosophical transactions of the Royal Society B: Biological sciences*, vol. 364, no. 1521, pp. 1211–1221, 2009.

[60] T. B. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20, Vancouver, BC, Canada: Curran Associates Inc., 2020, ISBN: 9781713829546.

[61] J. W. Rae, S. Borgeaud, T. Cai, *et al.*, *Scaling language models: Methods, analysis & insights from training gopher*, 2022. arXiv: 2112.11446 [cs.CL]. URL: https://arxiv.org/abs/2112.11446.

[62] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *Preprint arXiv:2001.08361*, 2020.

[63] J. Hoffmann, S. Borgeaud, A. Mensch, *et al.*, "Training compute-optimal large language models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22, New Orleans, LA, USA: Curran Associates Inc., 2024, ISBN: 9781713871088.

[64] P. Vogt, "The physical symbol grounding problem," *Cognitive Systems Research*, vol. 3, no. 3, pp. 429–457, 2002.

[65] N. Fishman and L. Hancox-Li, "Should attention be all we need? the epistemic and ethical implications of unification in machine learning," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22, Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 1516–1527, ISBN: 9781450393522. DOI: 10.1145/3531146.3533206. URL: https://doi.org/10.1145/3531146.3533206.

[66] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," pp. 66–71, 2018.

[67] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," pp. 770–778, 2016.

[69] J.-B. Alayrac, J. Donahue, P. Luc, *et al.*, "Flamingo: A visual language model for few-shot learning," *Preprint arXiv:2204.14198*, 2022.

[70] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[71] V. Sanh, A. Webson, C. Raffel, *et al.*, "Multitask prompted training enables zero-shot task generalization," in *International Conference on Learning Representations*, 2022. URL: https://openreview.net/forum?id=9Vrb9D0WI4.

[72] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *Preprint arXiv:2109.01652*, 2021.

[73] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-xl: Attentive language models beyond a fixed-length context," pp. 2978–2988, 2019.

[74] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning," pp. 1094–1100, 2020.

[75] S. Racanière, T. Weber, D. Reichert, *et al.*, "Imagination-augmented agents for deep reinforcement learning," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30, Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/9e82757e9a1c12cb710ad680db11f6f1-Paper.pdf.

[76] M. Chevalier-Boisvert, D. Bahdanau, S. Lahlou, L. Willems, C. Saharia, T. H. Nguyen, and Y. Bengio, "BabyAI: A platform to study the sample efficiency of grounded language learning," *Preprint arXiv:1810.08272*, 2018.

[77] S. Tunyasuvunakool, A. Muldal, Y. Doron, S. Liu, S. Bohez, J. Merel, T. Erez, T. Lillicrap, N. Heess, and Y. Tassa, "dm_control: Software and tasks for continuous control," *Software Impacts*, vol. 6, p. 100 022, 2020.

[78] C. Beattie, J. Z. Leibo, D. Teplyashin, *et al.*, *Deepmind lab*, 2016. arXiv: 1612.03801 [cs.AI]. URL: https://arxiv.org/abs/1612.03801.

[79] K. Zolna, A. Novikov, K. Konyushkova, C. Gulcehre, Z. Wang, Y. Aytar, M. Denil, N. de Freitas, and S. Reed, "Offline learning from demonstrations and unlabeled experience," *Preprint arXiv:2011.13885*, 2020.

[80] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, "The arcade learning environment: An evaluation platform for general agents," *Journal of Artificial Intelligence Research*, vol. 47, pp. 253–279, 2013.

[81] K. Cobbe, C. Hesse, J. Hilton, and J. Schulman, "Leveraging procedural generation to benchmark reinforcement learning," in *International conference on machine learning*, PMLR, 2020, pp. 2048–2056.

[82] W. Huang, I. Mordatch, and D. Pathak, "One policy to control them all: Shared modular policies for agent-agnostic control," in *International Conference on Machine Learning*, PMLR, 2020, pp. 4455–4464.

[83] A. Chowdhery, S. Narang, J. Devlin, *et al.*, "Palm: Scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, no. 1, Mar. 2024, ISSN: 1532-4435.

[84] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[85] M. Reid, Y. Yamada, and S. S. Gu, "Can Wikipedia help offline reinforcement learning?" *Preprint arXiv:2201.12122*, 2022.

[86] Q. Zheng, A. Zhang, and A. Grover, "Online decision transformer," *Preprint arXiv:2202.05607*, 2022.

[87] H. Furuta, Y. Matsuo, and S. S. Gu, "Generalized decision transformer for offline hindsight information matching," *arXiv preprint arXiv:2111.10364*, 2021.

[88] M. Janner, Q. Li, and S. Levine, "Offline reinforcement learning as one big sequence modeling problem," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[89] T. Chen, S. Saxena, L. Li, D. J. Fleet, and G. Hinton, "Pix2seq: A language modeling framework for object detection," in *International Conference on Learning Representations*, 2022. URL: https://openreview.net/forum?id=e42KbIw6Wb.

[90] A. Jaegle, S. Borgeaud, J.-B. Alayrac, *et al.*, "Perceiver IO: A general architecture for structured inputs & outputs," in *International Conference on Learning Representations*, 2022. URL: https://openreview.net/forum?id=fILj7WpI-g.

[91] V. Kurin, M. Igl, T. Rocktäschel, W. Boehmer, and S. Whiteson, "My body is a cage: The role of morphology in graph-based incompatible control," *Preprint arXiv:2010.01856*, 2020.

[92] C. Devin, A. Gupta, T. Darrell, P. Abbeel, and S. Levine, "Learning modular neural network policies for multi-task and multi-robot transfer," in *2017 IEEE international conference on robotics and automation (ICRA)*, IEEE, 2017, pp. 2169–2176.

[93] T. Chen, A. Murali, and A. Gupta, "Hardware conditioned policies for multi-robot transfer learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[94] S. Reed and N. De Freitas, "Neural programmer-interpreters," *International Conference on Learning Representations*, 2016.

[95] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[96] L. Kaiser, A. N. Gomez, N. Shazeer, A. Vaswani, N. Parmar, L. Jones, and J. Uszkoreit, "One model to learn them all," *Preprint arXiv:1706.05137*, 2017.

[97] J. Schmidhuber, "One big net for everything," *Preprint arXiv:1802.08864*, 2018.

[98] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "CTRL: A conditional transformer language model for controllable generation," *Preprint arXiv:1909.05858*, 2019.

[99] L. Espeholt, H. Soyer, R. Munos, *et al.*, "IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures," in *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, Eds., ser. Proceedings of Machine Learning Research, vol. 80, PMLR, Jul. 2018, pp. 1407–1416. URL: https://proceedings.mlr.press/v80/espeholt18a.html.

[100] H. F. Song, A. Abdolmaleki, J. T. Springenberg, *et al.*, "V-MPO: On-policy maximum a posteriori policy optimization for discrete and continuous control," in *International Conference on Learning Representations*, 2020. URL: https://openreview.net/forum?id=SylOlp4FvH.

[101] M. Hessel, H. Soyer, L. Espeholt, W. Czarnecki, S. Schmitt, and H. Van Hasselt, "Multi-task deep reinforcement learning with popart," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3796–3803.

[102] V. Mnih, K. Kavukcuoglu, D. Silver, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, 2015. URL: https://api.semanticscholar.org/CorpusID:205242740.

[103] Y. Tassa, Y. Doron, A. Muldal, *et al.*, *Deepmind control suite*, 2018. arXiv: 1801.00690 [cs.AI]. URL: https://arxiv.org/abs/1801.00690.

[104] J. Schrittwieser, I. Antonoglou, T. Hubert, *et al.*, "Mastering Atari, Go, chess and shogi by planning with a learned model," *Nature*, vol. 588, no. 7839, pp. 604–609, 2020.

[105] C. Gulcehre, Z. Wang, A. Novikov, *et al.*, "Rl unplugged: A suite of benchmarks for offline reinforcement learning," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20, Vancouver, BC, Canada: Curran Associates Inc., 2020, ISBN: 9781713829546.

[106] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Preprint arXiv:1810.04805*, 2018.

[107] J. M. Jumper, R. Evans, A. Pritzel, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, pp. 583–589, 2021. URL: https://api.semanticscholar.org/CorpusID:235959867.

[108] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," *arXiv preprint arXiv:2108.10904*, 2021.

[109] M. Chen, J. Tworek, H. Jun, *et al.*, *Evaluating large language models trained on code*, 2021. arXiv: 2107.03374 [cs.LG]. URL: https://arxiv.org/abs/2107.03374.

[110] Y. Li, D. Choi, J. Chung, *et al.*, "Competition-level code generation with alphacode," *Science*, vol. 378, no. 6624, pp. 1092–1097, 2022. DOI: 10.1126/science.abq1158. eprint: https://www.science.org/doi/pdf/10.1126/science.abq1158. URL: https://www.science.org/doi/abs/10.1126/science.abq1158.

[111] R. Nakano, J. Hilton, S. Balaji, *et al.*, *Webgpt: Browser-assisted question-answering with human feedback*, 2022. arXiv: 2112.09332 [cs.CL]. URL: https://arxiv.org/abs/2112.09332.

[112] R. Thoppilan, D. D. Freitas, J. Hall, *et al.*, *Lamda: Language models for dialog applications*, 2022. arXiv: 2201.08239 [cs.CL]. URL: https://arxiv.org/abs/2201.08239.

[113] V. Pratap, A. Sriram, P. Tomasello, A. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert, "Massively multilingual ASR: 50 languages, 1 model, 1 billion parameters," *Preprint arXiv:2007.03001*, 2020.

[114] R. Aharoni, M. Johnson, and O. Firat, "Massively multilingual neural machine translation," in *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)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3874–3884. DOI: 10.18653/v1/N19-1388. URL: https://aclanthology.org/N19-1388.

[115] R. Bommasani, D. A. Hudson, E. Adeli, *et al.*, *On the opportunities and risks of foundation models*, 2022. arXiv: 2108.07258 [cs.LG]. URL: https://arxiv.org/abs/2108.07258.

[116] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," *Preprint arXiv:2201.07207*, 2022.

[117] M. Ahn, A. Brohan, N. Brown, *et al.*, "Do as i can and not as i say: Grounding language in robotic affordances," in *arXiv preprint arXiv:2204.01691*, 2022.

[118] S. Li, X. Puig, C. Paxton, *et al.*, "Pre-trained language models for interactive decision-making," *Preprint arXiv:2202.01771*, 2022.

[119] S. Parisi, A. Rajeswaran, S. Purushwalkam, and A. Gupta, "The unsurprising effectiveness of pre-trained vision models for control," *Preprint arXiv:2203.03580*, 2022.

[120] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[121] S. Cabi, S. G. Colmenarejo, A. Novikov, *et al.*, *Scaling data-driven robotics with reward sketching and batch reinforcement learning*, 2020. arXiv: 1909.12200 [cs.RO]. URL: https://arxiv.org/abs/1909.12200.

[122] A. S. Chen, S. Nair, and C. Finn, "Learning generalizable robotic reward functions from "in-the-wild" human videos," *Preprint arXiv:2103.16817*, 2021.

[123] A. X. Lee, C. M. Devin, J. T. Springenberg, Y. Zhou, T. Lampe, A. Abdolmaleki, and K. Bousmalis, "How to spend your robot time: Bridging kickstarting and offline reinforcement learning for vision-based robotic manipulation," *Preprint arXiv:2205.03353*, 2022.

[124] Z. Wang, A. Novikov, K. Żołna, *et al.*, "Critic regularized regression," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20, Vancouver, BC, Canada: Curran Associates Inc., 2020, ISBN: 9781713829546.

[125] J. Harding, "Operationalising representation in natural language processing," *British Journal for the Philosophy of Science*, 2023.

[126] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," *Preprint arXiv:2005.00928*, 2020.

[127] B. Baker, I. Akkaya, P. Zhokhov, J. Huizinga, J. Tang, A. Ecoffet, B. Houghton, R. Sampedro, and J. Clune, "Video pretraining (vpt): Learning to act by watching unlabeled online videos," *Preprint arXiv::2206.11795*, 2022.

[128] J. Bruce, M. Dennis, A. Edwards, *et al.*, *Genie: Generative interactive environments*, 2024. arXiv: 2402.15391 [cs.LG]. URL: https://arxiv.org/abs/2402.15391.

[129] I. Beltagy, M. E. Peters, and A. Cohan, *Longformer: The long-document transformer*, 2020. arXiv: 2004.05150 [cs.CL]. URL: https://arxiv.org/abs/2004.05150.

[130] L. Weidinger, J. Uesato, M. Rauh, *et al.*, "Taxonomy of risks posed by language models," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22, Seoul, Republic of Korea: Association for Computing Machinery, 2022, pp. 214–229, ISBN: 9781450393522. DOI: 10.1145/3531146.3533088. URL: https://doi.org/10.1145/3531146.3533088.

[131] N. Bostrom, *Superintelligence*. Dunod, 2017.

[132] S. Russell, *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

[133] L. Ouyang, J. Wu, X. Jiang, *et al.*, "Training language models to follow instructions with human feedback," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22, New Orleans, LA, USA: Curran Associates Inc., 2024, ISBN: 9781713871088.

[134] Z. Kenton, T. Everitt, L. Weidinger, I. Gabriel, V. Mikulik, and G. Irving, "Alignment of language agents," *Preprint arXiv:2103.14659*, 2021.

[135] D. Amodei, C. Olah, J. Steinhardt, P. F. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," *Preprint arXiv:1606.06565*, 2016.

[136] C. Lu, J. Kay, and K. McKee, "Subverting machines, fluctuating identities: Re-learning human categorization," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1005–1015.

[137] G. C. Bowker and S. L. Star, *Sorting things out: Classification and its consequences*. MIT press, 2000.

[138] J. Butler, *Gender Trouble*. Routledge, 1990.

[139] H. R. Maturana and F. J. Varela, *Autopoiesis and Cognition: The Realization of the Living*. Springer Science & Business Media, 1991, vol. 42.

[140] W. I. Cho, J. Kim, J. Yang, and N. S. Kim, "Towards cross-lingual generalization of translation gender bias," FAccT '21, pp. 449–457, 2021. DOI: 10.1145/3442188.3445907.

[141] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "How we analyzed the compas recidivism algorithm," *ProPublica (5 2016)*, vol. 9, no. 1, 2016.

[142] M. J. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," *CoRR*, vol. abs/1711.05144, 2017. arXiv: 1711.05144.

[143] M. D. Storms, "Theories of sexual orientation.," *Journal of Personality and Social Psychology*, vol. 38, no. 5, p. 783, 1980.

[144] S. Ionescu, A. Hannák, and K. Joseph, "An agent-based model to evaluate interventions on online dating platforms to decrease racial homogamy," FAccT '21, pp. 412–423, 2021. DOI: 10.1145/3442188.3445904.

[145] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," Proceedings of Machine Learning Research, vol. 81, S. A. Friedler and C. Wilson, Eds., pp. 77–91, Feb. 2018.

[146] K. S. Ford, A. N. Patterson, and M. P. Johnston-Guerrero, "Monoracial normativity in university websites: Systematic erasure and selective reclassification of multiracial students.," *Journal of Diversity in Higher Education*, 2019.

[147] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," *arXiv preprint arXiv:1703.06856*, 2017.

[148] I. Kohler-Hausmann, "Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination," *Nw. UL Rev.*, vol. 113, p. 1163, 2018.

[149] A. Kasirzadeh and A. Smart, "The use and misuse of counterfactuals in ethical machine learning," FAccT '21, pp. 228–236, 2021.

[150] L. Hu and I. Kohler-Hausmann, "What's sex got to do with fair machine learning?" *arXiv preprint arXiv:2006.01770*, 2020.

[151] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[152] S. Cave, "The problem with intelligence: Its value-laden history and the future of AI," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 29–35.

[153] O. Keyes, Z. Hitzig, and M. Blell, "Truth from the machine: Artificial intelligence and the materialization of identity," *Interdisciplinary Science Reviews*, vol. 46, no. 1-2, pp. 158–175, 2021. DOI: 10.1080/03080188.2020.1840224.

[154] A. Hanna, E. Denton, A. Smart, and J. Smith-Loud, "Towards a critical race methodology in algorithmic fairness," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 501–512.

[155] E. Denton, A. Hanna, R. Amironesei, A. Smart, H. Nicole, and M. K. Scheuerman, "Bringing the people back in: Contesting benchmark machine learning datasets," *arXiv preprint arXiv:2007.07399*, 2020.

[156] C. Barabas, C. Doyle, J. Rubinovitz, and K. Dinakar, "Studying up: Reorienting the study of algorithmic fairness around issues of power," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 167–176.

[157] L. Hancox-Li and I. E. Kumar, "Epistemic values in feature importance methods: Lessons from feminist epistemology," in *proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 817–826.

[158] A. Birhane, "The Impossibility of Automating Ambiguity," *Artificial Life*, vol. 27, no. 1, pp. 44–61, Jun. 2021, ISSN: 1064-5462. DOI: 10.1162/artl_a_00336.

[159] D. Pfau and O. Vinyals, "Connecting generative adversarial networks and actor-critic methods," *arXiv preprint arXiv:1610.01945*, 2016.

[160] I. Grondman, L. Busoniu, G. A. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1291–1307, 2012.

[161] M. Miceli, M. Schuessler, and T. Yang, "Between subjectivity and imposition: Power dynamics in data annotation for computer vision," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1–25, 2020.

[162] A. Birhane, "Algorithmic injustice: A relational ethics approach," *Patterns*, vol. 2, no. 2, p. 100 205, 2021.

[163] S. Mhlambi, "From rationality to relationality: Ubuntu as an ethical and human rights framework for artificial intelligence governance," *Carr Center for Human Rights Policy Discussion Paper Series*, vol. 9, 2020.

[164] S. Hall and P. D. Gay, *Questions of Cultural Identity*. SAGE Publications, 1996.

[165] F. Garcıa-Peñalvo and A. Vázquez-Ingelmo, "What do we mean by genai? a systematic mapping of the evolution, trends, and techniques involved in generative ai," *International Journal of Interactive Multimedia and Artificial Intelligence*, 2023.

[166] E. M. Bender, T. Gebru, A. McMillan-Major, and S. 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*, 2021, pp. 610–623.

[167] C. Bird, E. Ungless, and A. Kasirzadeh, "Typology of risks of generative text-to-image models," in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023, pp. 396–410.

[168] F. Bianchi, P. Kalluri, E. Durmus, F. Ladhak, M. Cheng, D. Nozza, T. Hashimoto, D. Jurafsky, J. Zou, and A. Caliskan, "Easily accessible text-to-image generation amplifies demographic stereotypes at large scale," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 1493–1504.

[169] E. Ferrara, "Should chatgpt be biased? challenges and risks of bias in large language models," *arXiv preprint arXiv:2304.03738*, 2023.

[170] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[171] C. Vaccari and A. Chadwick, "Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news," *Social Media+ Society*, vol. 6, no. 1, p. 2 056 305 120 903 408, 2020.

[172] S. Monteith, T. Glenn, J. R. Geddes, P. C. Whybrow, E. Achtyes, and M. Bauer, "Artificial intelligence and increasing misinformation," *The British Journal of Psychiatry*, vol. 224, no. 2, pp. 33–35, 2024.

[173] A. Kasirzadeh and I. Gabriel, "In conversation with artificial intelligence: Aligning language models with human values," *Philosophy & Technology*, vol. 36, no. 2, pp. 1–24, 2023.

[174] R. McKinnon, "Gaslighting as epistemic injustice," *The Routledge handbook of epistemic injustice*, pp. 167–174, 2017.

[175] R. Tsosie, "Indigenous peoples and epistemic injustice: Science, ethics, and human rights," *Wash. L. Rev.*, vol. 87, p. 1133, 2012.

[176] J. Kay, A. Kasirzadeh, and S. Mohamed, "Epistemic injustice in generative AI," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, 2024, pp. 684–697.

[177] K. M. Davis, *Tainted Tap: Flint's Journey from Crisis to Recovery*. UNC Press Books, 2021.

[178] G. C. Spivak, "Can the subaltern speak?" In *Marxism and the Interpretation of Culture*, Macmillan Education, 1988, pp. 271–313.

[179] P. H. Collins, *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. Routledge, 2000.

[180] M. Fricker, *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press, 2007.

[181] N. Berenstain, "White feminist gaslighting," *Hypatia*, vol. 35, no. 4, pp. 733–758, 2020.

[182] K. W. Crenshaw, "Mapping the margins: Intersectionality, identity politics, and violence against women of color," in *The public nature of private violence*, Routledge, 2013, pp. 93–118.

[183] M. Fricker and K. Jenkins, "Epistemic injustice, ignorance, and trans experiences," in *The Routledge companion to feminist philosophy*, Routledge, 2017, pp. 268–278.

[184] J. Medina, "Misrecognition and epistemic injustice," *Feminist Philosophy Quarterly*, vol. 4, no. 4, 2018.

[185] R. Alvarado, "Ai as an epistemic technology," *Science and Engineering Ethics*, vol. 29, no. 5, p. 32, 2023.

[186] J. Byrnes and A. Spear, "Epistemic injustice and algorithmic epistemic injustice in healthcare," in *International Conference on Computer Ethics*, vol. 1, 2023.

[187] S. Glaberson, "The epistemic injustice of algorithmic family policing," *UC Irvine Law Review*, vol. 14, no. 2, 2024.

[188] G. Pozzi, "Automated opioid risk scores: A case for machine learning-induced epistemic injustice in healthcare," *Ethics and Information Technology*, vol. 25, no. 1, p. 3, 2023.

[189] J. Symons and R. Alvarado, "Epistemic injustice and data science technologies," *Synthese*, vol. 200, no. 2, p. 87, 2022.

[190] R. Wexler, "The odds of justice: Code of silence: How private companies hide flaws in the software that governments use to decide who goes to prison and who gets out," *Chance*, vol. 31, no. 3, pp. 67–72, 2018.

[191] G. Hull, "Dirty data labeled dirt cheap: Epistemic injustice in machine learning systems," *Ethics and Information Technology*, vol. 25, no. 3, p. 38, 2023.

[192] S. Milano and C. Prunkl, "Algorithmic profiling as a source of hermeneutical injustice," *Philosophical Studies*, 2024.

[193] A. L. Hoffmann, "Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse," *Information, Communication & Society*, vol. 22, no. 7, pp. 900–915, 2019.

[194] A. Kasirzadeh, "Algorithmic fairness and structural injustice: Insights from feminist political philosophy," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, pp. 349–356.

[195] E. Edenberg and A. Wood, "An epistemic lens on algorithmic fairness," in *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2023, pp. 1–10.

[196] M. De Proost and G. Pozzi, "Conversational artificial intelligence and the potential for epistemic injustice," *The American Journal of Bioethics*, vol. 23, no. 5, pp. 51–53, 2023.

[197] J. Brewster, L. Arvanitis, and M. Sadeghi, *The next great misinformation superspreader: How chatgpt could spread toxic misinformation at unprecedented scale*, https://www.newsguardtech.com/misinformation-monitor/jan-2023/, Jan. 2023.

[198] L. Arvanitis, M. Sadeghi, and J. Brewster, *Despite openai's promises, the company's new ai tool produces misinformation more frequently, and more persuasively, than its predecessor*, https://www.newsguardtech.com/misinformation-monitor/march-2023/, Mar. 2023.

[199] OHCHR, United Nations, *OHCHR assessment of human rights concerns in the xinjiang uyghur autonomous region, people's republic of china*, 2022.

[200] A. E. Marwick and R. Lewis, *Media manipulation and disinformation online*, https://www.posiel.com/wp-content/uploads/2016/08/Media-Manipulation-and-Disinformation-Online-1.pdf, 2017.

[201] J. Jaiswal, C. LoSchiavo, and D. C. Perlman, "Disinformation, misinformation and inequality-driven mistrust in the time of covid-19: Lessons unlearned from aids denialism," *AIDS and Behavior*, vol. 24, pp. 2776–2780, 2020.

[202] K. Nera, P. Bertin, and O. Klein, "Conspiracy theories as opportunistic attributions of power," *Current opinion in psychology*, vol. 47, p. 101 381, 2022.

[203] U. S. I. Committee, *Russian active measures campaigns and interference in the 2016 u.s. election.* Senate Report, 2020.

[204] K. J. Schiff, D. S. Schiff, and N. Bueno, "The liar's dividend: The impact of deepfakes and fake news on trust in political discourse," *SocArXiv*, 2023.

[205] N. Giansiracusa, "Deepfake deception: What to trust when seeing is no longer believing," *How Algorithms Create and Prevent Fake News: Exploring the Impacts of Social Media, Deepfakes, GPT-3, and More*, pp. 41–66, 2021.

[206] J. A. Goldstein, J. Chao, S. Grossman, A. Stamos, and M. Tomz, "How persuasive is AI-generated propaganda?" *PNAS nexus*, vol. 3, no. 2, pgae034, 2024.

[207] D. L. Byman, C. Gao, C. Meserole, and V. Subrahmanian, *Deepfakes and international conflict*, https://www.brookings.edu/articles/deepfakes-and-international-conflict/, Jan. 2023.

[208] H. S. Heroes, *State of deepfakes: Realities, threats, and impact*, https://www.homesecurityheroes.com/state-of-deepfakes, Dec. 2023.

[209] T. Lee and K. Koltai, *The folly of dall-e: How 4chan is abusing bing's new image model*, https://www.bellingcat.com/news/2023/10/06/the-folly-of-dall-e-how-4chan-is-abusing-bings-new-image-model/, Oct. 2023.

[210] R. Qadri, R. Shelby, C. L. Bennett, and E. Denton, "AI's regimes of representation: A community-centered study of text-to-image models in south asia," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 506–517.

[211] G. Pohlhaus, "Relational knowing and epistemic injustice: Toward a theory of willful hermeneutical ignorance," *Hypatia*, vol. 27, no. 4, pp. 715–735, 2012.

[212] J. Gurfinkel, *AI and the American smile*, https://medium.com/@socialcreature/ai-and-the-american-smile-76d23a0fbfaf, Mar. 2023.

[213] Z. Mengesha, C. Heldreth, M. Lahav, J. Sublewski, and E. Tuennerman, ""I don't think these devices are very culturally sensitive."—impact of automated speech recognition errors on African Americans," *Frontiers in Artificial Intelligence*, vol. 4, p. 169, 2021.

[214] D. Kazenwadel and C. V. Steinert, "How user language affects conflict fatality estimates in chatgpt," *arXiv preprint arXiv:2308.00072*, 2023.

[215] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," *arXiv preprint arXiv:1904.09751*, 2019.

[216] I. Augenstein, T. Baldwin, M. Cha, *et al.*, *Factuality challenges in the era of large language models*, 2023. arXiv: 2310.05189 [cs.CL]. URL: https://arxiv.org/abs/2310.05189.

[217] A. Abulimiti, C. Clavel, and J. Cassell, "How about kind of generating hedges using end-to-end neural models?" *arXiv preprint arXiv:2306.14696*, 2023.

[218] H. Zhang, S. Diao, Y. Lin, Y. R. Fung, Q. Lian, X. Wang, Y. Chen, H. Ji, and T. Zhang, "R-tuning: Teaching large language models to refuse unknown questions," *arXiv preprint arXiv:2311.09677*, 2023.

[219] K. Liu, S. Casper, D. Hadfield-Menell, and J. Andreas, "Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness?" *arXiv preprint arXiv:2312.03729*, 2023.

[220] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[221] A. Birhane, V. U. Prabhu, and E. Kahembwe, "Multimodal datasets: Misogyny, pornography, and malignant stereotypes," *arXiv preprint arXiv:2110.01963*, 2021.

[222] V. Prabhakaran, R. Qadri, and B. Hutchinson, "Cultural incongruencies in artificial intelligence," *arXiv preprint arXiv:2211.13069*, 2022.

[223] S. El-Sayed, C. Akbulut, A. McCroskery, *et al.*, *A mechanism-based approach to mitigating harms from persuasive generative ai*, 2024. arXiv: 2404.15058 [cs.CY]. URL: https://arxiv.org/abs/2404.15058.

[224] S. Ryan, "Epistemic environmentalism," *Journal of Philosophical Research*, vol. 43, pp. 97–112, 2018.

[225] G. Anderau, "Fake news and epistemic flooding," *Synthese*, vol. 202, no. 4, p. 106, 2023.

[226] K. Dotson, "Tracking epistemic violence, tracking practices of silencing," *Hypatia*, vol. 26, no. 2, pp. 236–257, 2011.

[227] K. Annesley, "Connecting epistemic injustice and justified belief in health-related conspiracies," *Ethics, Medicine and Public Health*, vol. 15, p. 100 545, 2020.

[228] B. Swire-Thompson, J. DeGutis, and D. Lazer, "Searching for the backfire effect: Measurement and design considerations," *Journal of applied research in memory and cognition*, vol. 9, no. 3, pp. 286–299, 2020.

[229] P. Bondy, "Argumentative injustice," *Informal Logic*, vol. 30, no. 3, 2010.

[230] P. Kalluri, "Don't ask if artificial intelligence is good or fair, ask how it shifts power," *Nature*, vol. 583, no. 7815, pp. 169–169, 2020.

[231] A. Birhane, W. Isaac, V. Prabhakaran, M. Diaz, M. C. Elish, I. Gabriel, and S. Mohamed, "Power to the people? opportunities and challenges for participatory AI," *Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–8, 2022.

[232] L. Groves, A. Peppin, A. Strait, and J. Brennan, "Going public: The role of public participation approaches in commercial ai labs," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 1162–1173.

[233] T. Kukutai and J. Taylor, *Indigenous data sovereignty: Toward an agenda*. ANU press, 2016.

[234] S. Huang and D. Siddarth, "Generative AI and the digital commons," *arXiv preprint arXiv:2303.11074*, 2023.

[235] A. Chan, H. Bradley, and N. Rajkumar, *Reclaiming the digital commons: A public data trust for training data*, 2023. arXiv: 2303.09001 [cs.CY].

[236] S. Fredman, D. Du Toit, M. Graham, K. Howson, R. Heeks, J.-P. van Belle, P. Mungai, and A. Osiki, "Thinking out of the box: Fair work for platform workers," *King's Law Journal*, vol. 31, no. 2, pp. 236–249, 2020.

[237] P. E. Agre, "Toward a critical technical practice: Lessons learned in trying to reform AI," in *Social science, technical systems, and cooperative work*, Psychology Press, 2014, pp. 131–157.

[238] D. Haraway, "Situated knowledges: The science question in feminism and the privilege of partial perspective," *Feminist studies*, vol. 14, no. 3, pp. 575–599, 1988.

[239] D. Ganguli, S. Huang, L. Lovitt, and D. Siddarth, *Collective constitutional AI: Aligning a language model with public input*, https://www.anthropic.com/news/collective-constitutional-ai-aligning-a-language-model-with-public-input, Oct. 2023.

[240] S. Gregory, "Fortify the truth: How to defend human rights in an age of deepfakes and generative AI," *Journal of Human Rights Practice*, vol. 15, no. 3, pp. 702–714, 2023.

[241] M. Schlichtkrull, Z. Guo, and A. Vlachos, "Averitec: A dataset for real-world claim verification with evidence from the web," *arXiv preprint arXiv:2305.13117*, 2023.

[242] X. Zeng, A. S. Abumansour, and A. Zubiaga, "Automated fact-checking: A survey," *Language and Linguistics Compass*, vol. 15, no. 10, e12438, 2021.

[243] T. Neumann, M. De-Arteaga, and S. Fazelpour, "Justice in misinformation detection systems: An analysis of algorithms, stakeholders, and potential harms," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1504–1515.

[244] E. Lett, E. Asabor, S. Beltrán, A. M. Cannon, and O. A. Arah, "Conceptualizing, contextualizing, and operationalizing race in quantitative health sciences research," *The Annals of Family Medicine*, vol. 20, no. 2, pp. 157–163, 2022.

[245] A. K. Heger, L. B. Marquis, M. Vorvoreanu, H. Wallach, and J. Wortman Vaughan, "Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–29, 2022.

[246] A. Grzankowski, "Real sparks of artificial intelligence and the importance of inner interpretability," *Inquiry*, pp. 1–27, 2024.

[247] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable AI: A brief survey on history, research areas, approaches and challenges," in *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*, Springer, 2019, pp. 563–574.

[248] W. Agnew, K. R. McKee, I. Gabriel, J. Kay, W. Isaac, A. S. Bergman, S. El-Sayed, and S. Mohamed, "Technologies of resistance to AI," *Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–13, 2023.

[249] L. Weidinger, M. Rauh, N. Marchal, *et al.*, *Sociotechnical safety evaluation of generative ai systems*, 2023. arXiv: 2310.11986 [cs.AI]. URL: https://arxiv.org/abs/2310.11986.

[250] D. Ganguli, L. Lovitt, J. Kernion, *et al.*, *Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned*, 2022. arXiv: 2209.07858 [cs.CL]. URL: https://arxiv.org/abs/2209.07858.

[251] U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis, "Generalization through memorization: Nearest neighbor language models," *arXiv preprint arXiv:1911.00172*, 2019.

[252] W. Freudenheim, I. Sekalala, and D. Zhu, *The ends of science*, https://endsofscience.org/the-ends-of-science/, Jul. 2023.

[253] E. M. Bender and A. Koller, "Climbing towards NLU: On meaning, form, and understanding in the age of data," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 5185–5198.

[254] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," *arXiv preprint arXiv:1909.08593*, 2019.

[255] Y. Bai, A. Jones, K. Ndousse, *et al.*, *Training a helpful and harmless assistant with reinforcement learning from human feedback*, 2022. arXiv: 2204.05862 [cs.CL]. URL: https://arxiv.org/abs/2204.05862.

[256] *HAPTIX: Hand Proprioception and Touch Interfaces | DARPA*. URL: https://www.darpa.mil/research/programs/hand-proprioception-and-touch-interfaces (visited on 02/07/2025).

[257] J.-A. Mbembé and L. Meintjes, "Necropolitics," *Public Culture*, vol. 15, no. 1, pp. 11–40, 2003, Publisher: Duke University Press, ISSN: 1527-8018. URL: https://muse.jhu.edu/pub/4/article/39984 (visited on 02/07/2025).

[258] *Donovan: Empowering the Public Sector with GenAI | Scale AI*. URL: https://scale.com/donovan (visited on 02/07/2025).

[259] *Palantir IR*. URL: https://investors.palantir.com/news-details/2024/Anthropic-and-Palantir-Partner-to-Bring-Claude-AI-Models-to-AWS-for-U.S.-Government-Intelligence-and-Defense-Operations/ (visited on 02/07/2025).

[260] *Anduril Partners with OpenAI to Advance U.S. Artificial Intelligence Leadership and Protect U.S. and Allied Forces | Anduril*. URL: https://www.anduril.com/article/

anduril-partners-with-openai-to-advance-u-s-artificial-intelligence-leadership-and-protect-u-s/ (visited on 02/07/2025).

[261] *Announcing The Stargate Project | OpenAI*. URL: https://openai.com/index/announcing-the-stargate-project/ (visited on 02/07/2025).

[262] "We are Google and Amazon workers. We condemn Project Nimbus," *The Guardian*, Oct. 2021, ISSN: 0261-3077. URL: https://www.theguardian.com/commentisfree/2021/oct/12/google-amazon-workers-condemn-project-nimbus-israeli-military-contract (visited on 02/07/2025).

[263] *Israeli authorities using facial recognition to entrench apartheid*, May 2023. URL: https://www.amnesty.org/en/latest/news/2023/05/israel-opt-israeli-authorities-are-using-facial-recognition-technology-to-entrench-apartheid/ (visited on 02/07/2025).

[264] L. Suchman, *The Algorithmically Accelerated Killing Machine*, Jan. 2024. URL: https://ainowinstitute.org/publication/the-algorithmically-accelerated-killing-machine (visited on 02/07/2025).

[265] H. Khlaaf, S. M. West, and M. Whittaker, *Mind the Gap: Foundation Models and the Covert Proliferation of Military Intelligence, Surveillance, and Targeting*, arXiv:2410.14831 [cs], Oct. 2024. DOI: 10.48550/arXiv.2410.14831. URL: http://arxiv.org/abs/2410.14831 (visited on 02/07/2025).

[266] L. Suchman, "Algorithmic warfare and the reinvention of accuracy," *Critical Studies on Security*, vol. 8, no. 2, pp. 175–187, May 2020, Publisher: Routledge _eprint: https://doi.org/10.1080/21624887.2020.1760587, ISSN: 2162-4887. DOI: 10.1080/21624887.2020.1760587. URL: https://doi.org/10.1080/21624887.2020.1760587 (visited on 01/22/2025).

[267] *Google fires 28 workers for protesting $1.2 billion Israel contract*. URL: https://www.nbcnews.com/news/us-news/google-fires-workers-protest-israel-contract-project-nimbus-rcna148333 (visited on 02/07/2025).

[268] S. Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, 2019, ISBN: 9781610395700. URL: https://books.google.com/books?id=lRqrDQAAQBAJ.

[269] N. Tomasev, K. R. McKee, J. Kay, and S. Mohamed, "Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities," en, in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, Virtual Event USA: ACM, Jul. 2021, pp. 254–265, ISBN: 978-1-4503-8473-5. DOI: 10.1145/3461702.3462540. URL: https://dl.acm.org/doi/10.1145/3461702.3462540 (visited on 01/23/2025).

[270] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka, and B. Y. Zhao, *Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models*, arXiv:2302.04222 [cs], Aug. 2023. DOI: 10.48550/arXiv.2302.04222. URL: http://arxiv.org/abs/2302.04222 (visited on 02/07/2025).

[271] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng, and B. Y. Zhao, *Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models*, arXiv:2310.13828 [cs], Apr. 2024. DOI: 10.48550/arXiv.2310.13828. URL: http://arxiv.org/abs/2310.13828 (visited on 02/03/2025).

[272] D. Spade, *Mutual Aid: Building Solidarity During This Crisis (and the Next)*. Verso Books, 2020.

[273] D. McQuillan, "People's Councils," in *Resisting AI*, ser. An Anti-fascist Approach to Artificial Intelligence, 1st ed., Bristol University Press, 2022, pp. 119–134, ISBN: 978-1-5292-1349-2. DOI: 10.2307/j.ctv2rcnp21.10. URL: https://www.jstor.org/stable/j.ctv2rcnp21.10 (visited on 02/07/2025).

[274] L. Parisi, "The alien subject of AI," *Subjectivity*, vol. 12, pp. 27–48, 2019.

[275] D. McQuillan, "Post-machinic Learning," in *Resisting AI*, ser. An Anti-fascist Approach to Artificial Intelligence, 1st ed., Bristol University Press, 2022, pp. 104–118, ISBN: 978-1-5292-1349-2. DOI: 10.2307/j.ctv2rcnp21.9. URL: https://www.jstor.org/stable/j.ctv2rcnp21.9 (visited on 02/07/2025).

[276] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* '19, Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 220–229, ISBN: 9781450361255. DOI: 10.1145/3287560.3287596. URL: https://doi.org/10.1145/3287560.3287596.