Perceptions of Agency: Untangling the knotty web of AI

by

Ziv Epstein

B.A., Pomona College (2017)

Submitted to the Program of Media Arts and Sciences, School of Architecture and Planning
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2019

Signature redacted

Author .....
Program of Media Arts and Sciences, School of Architecture and Planning
May 10, 2019

Signature redacted

Certified by.....
Iyad Rahwan
Associate Professor
Thesis Supervisor

Signature redacted

Accepted by .....
Tod Machover
Academic Head, Program in Media Arts and Sciences

Perceptions of Agency: Untangling the knotty web of AI

by

Ziv Epstein

Submitted to the Program of Media Arts and Sciences, School of Architecture and Planning
on May 10, 2019, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

Artificial intelligence systems (AI) have become a ubiquitous part of modern life. Yet their complexity has prevented a concrete conceptualization that correctly map the web of human actors and computational processes involved in AI. This opaque representation of AI poses questions for accountability and governance, such as who is responsible when an AI makes a moral transgression? This thesis takes a discursive and empirical approach to reifying AI as a specific network of human actors with real world outcomes. It explores the phenomenon of anthropormophization, by which AI is endowed with human-like characteristics, and shows how the extent to which a AI system is anthropomorphized can affect the attribution of responsibility to human actors. This thesis does not offer a normative suggestion for whom society should blame when AI make moral transgressions, but rather offers a view into human folk intuitions.

Thesis Supervisor: Iyad Rahwan
Title: Associate Professor

This masters thesis has been examined by a Committee as follows:

Dr. Iyad Rahwan ..... Signature redacted .....

Chairman, Thesis Committee

Associate Professor of Media Arts and Sciences

MIT

Dr. Andrew Lippman ..... Signature redacted .....

Member, Thesis Committee

Senior Research Scientist and Associate Director of the MIT Media Lab

MIT

Dr. Nick Seaver ..... Signature redacted .....

Member, Thesis Committee

Assistant Professor of Anthropology

Tufts University

Acknowledgments

Just as AI systems are complex webs of humans actors, this thesis too would not have been possible without a vital network of collaborators, intellectual mentors, friends and loved ones.

I would like to thank my advisor Iyad Rahwan for creating a space for me to explore for two mind-bending years. I would also like to thank my thesis readers, Andy Lippman and Nick Seaver, for their direction. A special thanks to Nick for opening up the world of cultural anthropology to me and for being patient as the thesis morphed and evolved into its final form.

I would like to thank Sydney Levine who kindly offered her time and expertise towards making the thesis what it is. Without her wisdom on experimental methods and constant encouragement, this thesis would certainly not exist. I would also like to thank the mystic Matt Groh, whose constant questioning, support and collaboration were central to the culmination of these ideas. The time spent vision-questing in the Nova Scotia wilderness and training deep learning models at the wee hours of morning were some of the most fun and formative times of this work. I would also like to thank the other collaborators involved in the technical and experimental aspects this work, which includes Dave Rand, Nick Obradovich, Manuel Cebrian, Abhimanyu Dubey and Niccolo Pescetelli. In addition, I would like to thank the *crowd*, the 5,634 people who provided the images to DeepAngel that we used to train AI Spirits, and the 3,927 Mechanical Turk workers who participated in 9 studies, for their labor.

I would also like to thank the intellectual titans that seeded the ideas for this thesis. A special thanks to Cathy O'Neil who made me realize that how we talk about and conceptualize our AI systems has important consequences for their ethical use. I would also like thank Joi Ito and Jonathan Zittrain, who taught the *Ethics and Governance of AI* class which served as the incubation site for many of these ideas. This thesis would not have been possible without the constant encouragement, camaraderie, and adversarial probing of Blakeley Payne Hoffman, who has been a massive intellectual inspiration on all things design, media, AI, and ethics from day one. I would also like

to thank Pinar Yanardag and Richard Kim, whose experiments on copyright and the perception of AI generated-artwork directly informed many of the ideas found here. I'm grateful for the feedback and wisdom of other academic giants, including Ethan Zuckerman, Janelle Shane, Edmond Awad and Sandy Pentland.

This thesis emerged directly from the many conversations and interactions with the amazing friends I've made along the way. I would like to thank Judy Shen, May Alhazzani, Anna Chung, Josh Hirschfeld-Kroen, Kalli Retzepi, Agnes Cameron, Pip Mothersill, Dima Smirnov, Michiel Bakker, Nick Stagnaro, Adam Bear, Erez Yoeli, Gordon Kraft-Todd and Antonio Alonso-Arechar for their constant feedback, support, ideas, and vibes.

I would like to thank Amna Carreiro for her help and support navigating the Scalable Cooperation group and the Media Lab in general.

Finally, I would like to thank my family for their love and patience during this process. Thank you Micah for your intense wisdom on all things design, and being a constant source of inspiration and support. Thank you Charlie for being a shoulder to lean on, and for always taking the edge off life. Thank you Mom and Dad for helping triage details as I was writing this thesis and giving me the strength to make it to the finish line.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

New troves of high resolution data about human behavior and practices, coupled with powerful new information processing systems, have resulted in artificial intelligence (AI) making decisions and exerting influence in nearly every aspect of our lives. AI now assists us in making decisions as simple as what movie to watch next or which route is optimal for avoiding traffic. But AI is now involved in more complex, high-stakes decision making. AI systems are being deployed in the criminal justice system to assist with criminal sentencing. Newsfeed algorithms on social media determine what content we see, and we are starting to share the roads with AI-powered autonomous vehicles. Thus, understanding how these systems work, and how we should interact with them, is of principal importance for citizens, scientists and the public at large.

One such domain where there has been little empirical work is the context of art, where humans and machines collaborate to create cultural artifacts that are evaluated by subjective and cultural means. While the practices of creating art have always been deeply entwined with exploring the affordances of new technologies, the story of AI Art is particularly interesting due to the emergence of a new technology: the GAN. A generative adversarial network (GAN), created by Goodfellow et al, adapts the conventional supervised learning task to a zero-sum game between a generator (which attempts to produce images undifferentiable from the training data) and a discriminator (which attempts to discern which images are generated, contrasted with those from the training data) [42]. What results from this algorithmic dance is a

model that can produce novel instances of data. The advent of the GAN represents a fundamental shift in deep learning from prediction to creation, and as such has been adopted by the art community in droves. The GAN has become so popular as a tool for AI artists, as to have spawned a new modern art trend onto itself dubbed by Fançois Chollet as "GANism" [81, 26]. Recent events surrounding this new art form have raised many fundamental questions about the ethics of artificial intelligence.

Figure 1-1: Edmond De Belamy

On October 25, 2019, a portrait generated by a GAN sold at Christies art auction for $432,500 (see Figure 1-1). Since Christie's initial estimate for the piece by the unknown Parisian art collective Obvious was $10,000, its sale for over 40 times this expectation shocked the art world. Marketed by Christies as "the first portrait generated by an algorithm to come up for auction," the painting entitled "Edmond De Belamy" indeed struck a chord in society about the nature of authorship and artificial intelligence [29]. But the reality of the painting's creation is not as simple

as Christies purports. As shown in the last row of table 1, the process that resulted in the creation of Edmond de Belamy involves many people, from the Renaissance masters who painted the images that constitute the training data set to the Machine Learning researchers who published the algorithms used to construct it. In particular, the code used to generate the painting was written by the 19-year-old technologist and AI artist Robbie Barat in a MIT-licensed GitHub repository. As he puts, Obvious "almost immediately started producing work identical to the outputs of the pre-trained portrait and landscape networks" he had put online. Many influential AI artists agree with Barat's assertion that his work had been used without proper attribution. Mario Klingeman told the magazine Verge that "You could argue that probably 90 percent of the actual 'work' was done by [Barrat]" [109]. As Obvious took the code, trained the network, printed it and sent it to Christies, they are a part of the human actors responsible for the work. But how authorship and attribution of responsibility works in this case in unclear.

Figure 1-2: Adapted Edmond de Belamy [18]

Another fascinating layer to the Edmond de Belamy story is how the painting

was marketed. The early press materials of Obvious explicitly reference the AI as as the artist. In a press release in January, they told reporters that "an artificial intelligence managed to create art" which underpinned their motto that "creativity isn't only for humans." This marketing approach was the spark to the fuel of AI hype that resulted in the painting's enormous sale, and has raised provocative questions about the nature of AI systems. In a tweet a week after the sale, Waltz Binaire raises the point that another AI system, the "Content Based Fill" functionality built into Photoshop, is already a ubiquitous part of digital art, yet this AI is not put on the same pedestal as the Generative Adversarial Network used to create Edmond de Belamy (see Figure 1-2) [18]. This cunning marketing scheme suggests that the way we talk about AI, and in the particular the extent to which we endow our systems with agency, can have important consequences in the ethics of governance of AI. But what are the consequences of such a scheme? Did framing the AI in a particular way change the way people interpreted it?

This thesis uses the fascinating case of the Edmond de Belamy painting as a jumpoff point to explore these questions of distributed actors, allocation of credit and anthropormorphization in AI systems at large. In Chapter 1, I propose a conceptual framework for AI systems which is designed to generalize to many AI contexts, and to provide a structured way of thinking about how agency flows within them. In Chapter 2, I operationalize this framework into a critical art practice to explore how this framework fits with a real world community and data pipeline. In Chapter 3, I empirically investigate how people reason about ethical dilemmas involving AI, when cast in this particular light. Through a series of interrogations that spans disciplines and literatures, I hope to create a broader discussion about the implications of how we reify AI and to provide a suite of discursive and scientific tools to examine these questions in detail.

Chapter 2

The Knotty Web of AI

AI systems exhibit a complexity that makes it challenging (and impossible in some cases) for individuals to understand them and interact with them in a socially optimal manner. There are many reasons for this complexity, but I will discuss three here.

First, AI systems are not discrete objects that can be isolated and put under a microscope. Rather, they are "heterogeneous and diffuse and sociotechnical systems" that span many human actors and computational processes [96]. For example, consider the many people and processes responsible for Facebook's newsfeed algorithm. There are Facebook's software engineers, who program algorithms built on vast codebases collectively maintained by thousands of people. These algorithms are trained on high resolution data from the behavior of Facebook users – such as which videos they watch and what links they click on. These behaviors are in turn affected by Facebook's algorithms, which are pitted in a recursive feedback loop. Thus Facebook's newsfeed, distributed across thousands of servers, actively adapting to the behavior of millions of people, and coordinated through the actions of hundreds of programmers, resembles a government more than an algorithm like Bubblesort [96].

This algorithmic process of learning, heralded by powerful machine learning algorithms and petabytes of human behavioral data, yields a second layer of complexity. Modern machine learning methods of neural networks exhibit the properties of a black box: even the scientists and theorists who focus on these systems do not understand how they work, or why their performance is so good. These black box methodologies

make it difficult to audit or understand the behavior of these systems, which in turn results in an uptick in unexpected behavior [23, 34]. This can cause emergent behaviors like when Amazon's Alexa, unprompted, started mysteriously laughing. It can also result in more malicious behaviors like algorithmic bias, which has garnered large attention within both academia and industry [82, 38]. A recent study revealed that many of the top commercial facial recognition algorithms misgender dark, female faces at significantly higher rates than their lighter, male counterpart [21]. And in the criminal justice algorithm discussed above, ProPublica journalists discovered that the algorithm also misclassifies black defendants at higher rates than white defendants [9]. But the algorithm in question, COMPAS, is not a neural network, but rather a simple and interpretable regression model. What makes COMPAS a black box is not its algorithmic architecture but rather the fact that North Pointe, the company who created it, has gone to extreme lengths to ensure privileged access in order to protect their secret sauce (under the hood, COMPAS is a logistic regression, a comparatively simple model). Indeed most of the ubiquitous AI systems today are available not as fully accessible code but rather as priced API endpoints which treat the underlying model as an input/output black box. Thus both the corporate culture of treating AI systems as API endpoints, and the neural architecture that composes many models today, are both responsible for them to be black boxes, shrouded in mystery.

A final factor that contributes to the complexity is a "terminological anxiety" that pervades conversations of AI systems [96]. At her 2019 keynote at the Workshop on Ethical, Social and Governance issues in AI at NeurIPS, Hannah Wallach repeatedly emphasized how the media misuses the world "algorithm" in reference to AI systems [4]. Wallach is correct that in many applied settings, the concept of an algorithm suffers from a definitional ambiguity. In the academy, this manifests as a diverse conceptualization of algorithms, from a technological view [28] to an epistemological view [53]. Thus the very concept of an algorithm at once is hard to pin down and siloed by discipline, which in turn makes it easy for practitioners and academics to talk past one another [14, 96].

These layers situate AI as a complicated knot at the intersection of systems

thinking, technopolitics and modernity at large. Paola Antonelli, Neri Oxman and Kevin Slavin coined the term "Knotty Objects" to refer to objects for which "conception, design, manufacturing, use and misuse are non-linear [and] non-discrete [by] entangling practices, processes and politics" [10]. The invocation of the knotty-ness of AI here makes clear the fact that inter-(and anti-)disciplinary explorations are necessary to adequately conceptualize AI and its impacts. For indeed these layers of complexity not only form the diffuse, black-boxed, and ambiguous web of agents that constitute AI systems, but also pose daunting ethical questions of governance and accountability. In particular, how should responsibility be distributed across this web when the AI system makes a moral transgression? The diffuse nature of AI systems makes it exceedingly difficult to assign culpability to a single programmer, company executive or data provider. The black-boxed nature of AI systems makes it increasingly easy to blame the "inherent weirdness of machine learning" instead of the involved human actors. And moreover the modern complicated nature of AI systems makes it difficult to port existing norms and heuristics from social contexts into this novel algorithmic domain. Emerging interdisciplinary initiatives and research communities, such as the AI Now Institute, the Partnership on AI, and the Workshop on Fairness, Accountability and Transparency in Machine Learning [2] and the FAT* conference [1] are tackling these problems head on. Yet these initiatives and communities are principally interested in developing new measures of algorithmic fairness, and new models that mitigate these concerns, while less emphasis has been placed on either studying how people port over existing knowledge to make sense of AI systems, or creating unified frameworks by which to assign responsibility to human actors in the face of a moral transgression [56].

This question of responsibility mediated by technology is not new and has deep roots in STS and anthropology. Scholars have explored how perceptions of agency, which often tracks with folk notions of responsibility and is considered an essential pillar of moral reasoning, can be considered within a broader sociotechnical framework [117, 7, 44]. Vygotsky [111, 110] and Bateseon [15] argue that agency cannot be considered as an atomic characteristic of individuals, but rather is situated amongst

groups and necessary requires "mediational means such as tool and language" [7]. These mediational means are particularly salient for instance where emerging technologies impact the way we consider human actors. Although a sociotechnical definition of agency that accounts for the nuance of AI systems is beyond the scope of this work, I follow the bulk of the Moral Psychology literature, which suggests that the perception of agency is a crucial aspect of moral reasoning [43, 44]. [1]

Despite these interdisciplinary ventures to unpack issues of fairness, accountability, transparency and agency in these sociotechnical systems, their complexity has made it difficult to apply knowledge such as moral heuristics, scholarship, and laws to this new domain. On an evolutionary timescale, we as a species have cultivated social heuristics that shape our moral intuitions [50, 91]. While these heuristics enable cooperation, punishment and other prosocial behavior in a socially optimal equilibrium, they are honed in linear and discrete interagent environments (such as 2-player interactions like the prisoners dilemma). Thus, these intuitions break down in the face of complex and abstract dilemmas like those posed by AI systems. Our legal system is similar rooted in similar moral frameworks of discrete interagent interactions. This concept is articulated well by Selsbt et al [97] as five abstraction traps that occur when trying to conceptualize technical systems within a policy context. These traps are the framing trap, the portability trap, the formalism trap, the ripple effect trap, and the solutionism trap. These traps suggest that when applying our legal frameworks and moral intuitions to situations regarding AI, we are fighting an uphill battle.

2.1 Mapping the contours of AI Systems

While defining such traps provides insight into behaviors to avoid when attempting to consider moral transgressions of AI systems, there is no guarantee that new traps will not emerge as AI systems gain access to more user data and are deployed in

[1] While the bulk of this literature focuses on the Euro-American setting, there is a growing literature (pioneered by Ara Norenzayan and Aiyana Willard) that studies the cross-cultural aspects of mind perceptions [80, 119, 40]. This body of work investigates religiosity as it relates to mind perception, towards understanding the culturally stable phenomenon of anthropomorphization and belief in god.

more high-stakes social situations. Thus, this thesis attempts a conceptual framework for understanding AI systems, and explores how human moral intuitions map onto this conceptualization. Thus, it is both normative and descriptive in various ways. This framework offers normative guidance by providing policy-makers and ethicists new and coherent ways to consider AI systems and their constitute actors. However, this framework does not claim an actionable positivist set of ethics. Instead, I use empirical methods to describe how this framework intersects with how people actually reason about AI systems.

This framework traces Striphas' concept of algorithmic culture [104], which explores the dynamics by which algorithmic processes now conduct the "sorting, classifying and hierarchizing of people, places, and ideas." Striphas charts the "diffuse web of human actors and information processing systems [that] collectively co-construct algorithmic culture" and offers a framework by which to consider this web. This framework is based on three keywords that exert influence on culture in the face of the delegation of the work of culture to algorithmic processes: information, algorithm and crowd. Here, I build upon Striphas' meditations on these three words, towards constructing the actor network of human stakeholders that constitute an AI system.

2.1.1 Term 1: Information

Striphas charts the transformation of the concept of information from "utilitarian things like automatic door openers and thermostats" to "cultural objects, practices and preferences [as] a corpus of data." While the substrate of information, the 0s and 1s of data, remains constant, the content of this information begins to resemble the behavioral culture of the internet age. Indeed each day 500 million tweets are sent, 5 billion youtube videos are watched and 4.75 billion pieces of content are shared on Facebook. Mobile phones equipped with GPS track the real time location of 2.5 billion people, while credit card transactions follow consumers [67]. And in each case, there exist databases that store the data containing these behaviors, as well as personal metadata such as location, income, gender, and consumer preferences. This vast informational infrastructure means that an increasing portion of human cultural

practices and behaviors are quantified, catalogued and operationalized.

2.1.2 Term 2: Algorithm

These vast troves of human behavioral data have paved the way for powerful algorithms to enter into the cultural domain. Supervised machine learning algorithms, which constitute the workhorse of most modern AI systems, use this large amount of cultural information to learn representations of human cultural practices and produce predictions conditioned on priors. Some of these learning tasks, like which ad to serve to an online newspaper reader to maximize the probability they purchase the corresponding advertised product, or which piece of content to serve to a social media user to maximize engagement, are tractable problems with highly performant models. Other tasks like producing "aesthetically optimal" images remain intractable and reductive. Within the machine learning community, supervised machine learning refers to domains where learning occurs from a input/output pairs. The labelled output represents ground truth, and the task is to learn a optimal mapping between inputs and outputs. Here, we enact the notion of supervision differently by explicitly invoking the proximity of a human supervisor. Indeed many instances of machine learning across academia and industry involve close interaction between humans and AI (often dubbed "Human-in-the-loop"). Since here we are interested in the human aspects of machine learning, we will thus focus on those systems with the a human supervisor in the loop.

2.1.3 Term 3: Crowd

In the context of these supervised machine learning algorithms, Striphas' notion of the crowd can be enacted into two ways. The first follows from the modern concept of information above. That is, the crowd is the collective set of millions of people who create the behavioral data on which these algorithms are trained. By creating small amounts of data that indirectly shape these important algorithms, users are "denied... the the possibility of 'full participation', while still granted... a modicum

of effect or influence" [104]. By providing the ability for individuals to "vote" on the behavior of these algorithms, these AI systems provide an algorithmic social contract that keeps "society in the loop" [89]. However, aggregating this individual level data in a way that avoids tyranny of the majority, or latent human biases, remains an open research area [79, 99]. The second enactment of the crowd involves the set of people explicitly involved in the production and deployment of an AI system. This involves the technologists who develop the core algorithms and other computational infrastructure upon which AI systems are built. It also involves the practitioners who take this infrastructure and create AI systems by training the algorithms on user data and deploying them. In addition, it involves curators who receive the predictions for the model and decide whether or not to do as it says [27, 9]. Each of these actors plays an important role in the deployment of a AI system build on a supervised machine learning model. Yet like the crowd that constitutes the training data for the model, none of them is "fully conscious" of the "complex organization" [104, 120]. So then how do these crowds interrelate? And how should responsibility be distributed across these actors when an AI system makes a moral transgression?

2.2 An actor network of human stakeholders

Inspired by actor-network theory and Zhu and Harrell's AI Hermeneutic Network [125], I chart an actor network of human stakeholders who, alongside a suite of computational processes, compose the AI systems that involve the supervised machine learning models discussed above (see Figure 2-1).

In this actor network, the crowd discussed above collectively contribute behavioral data for algorithm training. These individuals submit data to a platform or other data repository. A technologist theorizes and/or programs a new machine learning algorithm or codebase. Yet without an application or deployed setting, this technologist's work remains abstract. A practitioner takes the work of the technologist, and the data from the crowd, to create a trained machine learning model within an application domain that can be accessed in real time. To make a prediction (at test time), this system

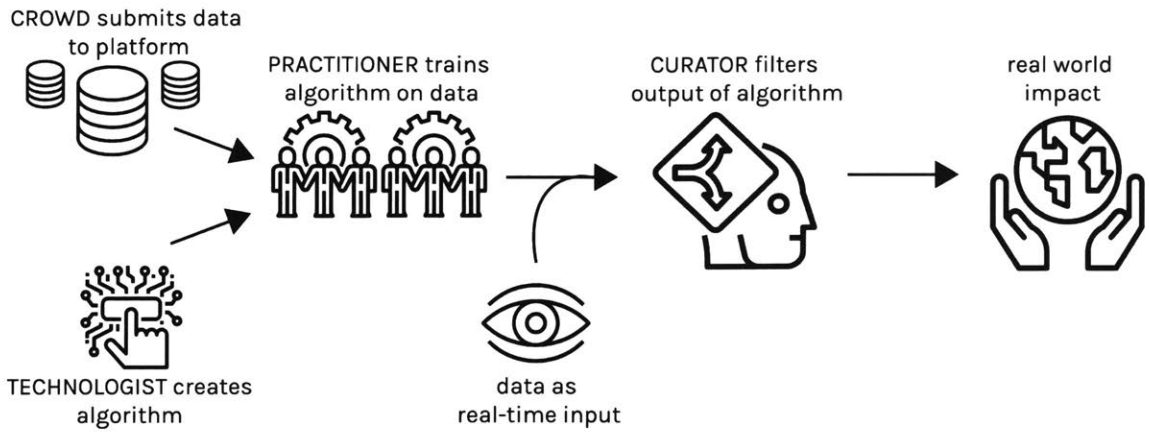


Figure 2-1: An actor network for supervised machine learning systems.

takes in a novel input (here referred to as the image) and produces a prediction. This prediction is then passed off to a human curator, who evaluates the prediction before deploying it in the world. Examples of each of these roles for many real world contexts are listed in Table 2.2. This actor network was inspired by discussing the process of creating AI art with artists from the community, as is discussed further in Section 3. Most of the literature on human in the loop systems, as well as the actor network I have proposed here, assume that humans maintain agency while cooperating with each other and machines in large complex sociotechnical systems. There is a phenomenon, however, called the Moral Crumble Zone, which posits that human actors may act as sponges to absorb blame and responsibility for the failings of large systems in which they may not have much agency [32]. For example, consider the airline representative at the desk of a cancelled flight, who bears the anger of frustrated travellers, despite having no agency over the cancellation itself, or any power to fix it. We see this occur with AI systems as well, such as the mission specialists in self-driving cars, who are responsible for "watching the autonomous software to ensure no one gets hurt" [73]. In this framework, the moral crumble zone is primarily manifested as the "curator," who stands as a filter between algorithmic decision making and real world impact and thus be prone to soaking up moral outrage from the mistakes of AI systems. However, this theory has not yet been empirically tested, and it is unclear how the moral crumble zone interacts with other competing actions within AI systems (such as the crowd

AI System	Crowd	technologist	practitioner	curator	real world impact
Self-driving car	Drivers who log Thousands of hours of manual driving data	Telsa engineers such as Andrej Karpathy, inventors of new Reinforcement Learning and Machine Learning methods	Tesla and Waymo engineers	Human operator in car (mission specialist)	People get hit and die
Risk recidivism Algorithm (COMPAS)	Criminals, whose demographic data and outcomes are provided for training	Logistic regression	Northepoint	Judge	Which criminals get to re-enter society and which stay in prison
Newsfeed	Social Media Users	Faceook engineers like Lars Backstrom, inventors of new machine learning methods	Facebook software engineers	content moderator	Echo chambers and polarization
Deep Angel [45]	Annotated masks (for Mask-rCNN) and images (for generative inpainting)	Mask-rCNN and generative inpainting, so ML academics	The Scalable Cooperation Team	None	Mass media manipulation through the omission and censoring of content
GAN art	A large number of digitized paintings	GANs (Ian Goodfellow) and programmers who make github repos (Robbie Barat)	Obvious	Obvious	Painting sells to Christy for 432k
Predictive Diagnosis tool for healthcare	patients annotated for disease as ground truth	machine learning researchers	consultant who integrates ML systems into healthcare pipelines	doctor	possible biases in predictions, and improved efficiency of healthcare

Table 2.2.1: Examples of the human stakeholders in various contexts where AI is deployed in real world setting.

providing systematically biased data, or the practitioner creating a flawed model).

This simplified model contains many of the salient and core features of many AI systems, and serves as model that connects many different notions of artificial intelligence. In particular, it introduces a group of individuals who are each causally related to the outcome of the AI system. That is, following the moral psychologies use of the counterfactual reasoning to construct causal graphs, if each of these actors did not act as they did, then the final outcome what not have occurred [86, 64].

However, it comes with many caveats. First of all, it neglects the many other humans involved in AI systems Second, it reductively assumes a fixed and discrete causal structure with a few number of causal arrows. Indeed in actuality, all of these actors and processes influence each other in subtle and important ways. For example, consider the newsfeed example discussed above. For a newsfeed, there is a feedback loop in the output of the algorithm itself (i.e. which content is shown to the user) that alters the behavior of the users. Third, it does not rigorously account for situations in which the same human occupies many of the same roles. Finally, it must be emphasized that this model is only one possible abstraction created for explaining a limited set of AI situations. Since this thesis does not provide a strong epistemological basis to validate or test this model, its remains future work to see how it compares with other possible models.

2.2.1 Anthropomorphism mediates the actor network

Humans, however, use more than just a causal graph of actions and results to reason about moral quandaries. The thrust of the moral psychology literature suggests that our perceptions about the minds of those involved, which manifest as intentions, beliefs and values, (as Gray, Young and Waytz say: "Mind perception is the essence of morality" [44]) meaningfully vary across individuals, and shape our moral judgments [114, 33, 43, 44]. This well-established effect is compounded by the phenomenon in both academia and the general public where AI is endowed with unsubstantiated anthropomorphic characteristics. As showcased by Lipton and Steinhardt's "Troubling Trends in Machine Learning Scholarship" [70], many tasks and techniques in the

literature are specified using the same language as they would for a human, such as reading comprehension [49], music composition [78], curiosity [94], fear [69], thought vectors [63] and consciousness priors [16].

While these anthropomorphizations can be useful for providing a useful analogy or commutating inspiration, they are not neutral. Rather, it has been shown that the anthropomorphization of computational process can alter human moral judgements in relations to these processes. Through a series of experiments involving an unavoidable crash in a driving simulator with cars of varying complexity (i.e. a normal car versus a self driving car versus an anthropomorphized self driving car with a human voice, and name) Waytz and colleagues demonstrated that increases in the anthropomophization of a car predicts behavioral, psychological and physiological measures of trust in the car [115]. While they mostly focused on the psychological construct of trust, they also found that anthropomorphization affected attributions of responsibility and punishment for the car's mistakes, which is consistent with the established relationship between agency and perceived responsibility [33, 115]. Thus, it is clear that the extent to which computational systems are anthropomorphized matters, and may bear upon how we distribute responsibility to the human actors that constitute AI systems. Waytz, Heafner and Elpey even ask about the blame of the various actors involved, such as the participant themselves, the car, the car designer, and the the car company. However, they collapse the last three measures into a single measure of "blame for vehicle" which in turn neglects the complex web of actors involved in a self-driving car. In this work, we seek to extend the insights from this line of work, while accounting explicitly for this web of actors.

2.3 Towards an empirical investigation of machine agency

As we begin to encounter more social dilemmas that involve AI, we need to new tools, techniques and conceptualizations to handle the increasing complexity of these

dilemmas. In this chapter, I have proposed a new conceptual framework for thinking about AI systems. It is designed to follow the underlying technical practices involved in architecting and deploying AI systems, and thus generalizes to many different AI contexts. While different AI contexts, from recommender systems to self-driving cars have their own set of moral stakes and stakeholders, this framework points to a hopeful universality of machine ethics.

This framework is designed to particularly add clarity to instances where the knotty-ness of AI systems make it hard to reason about issues of credit, blame and responsibility. Towards that end, this thesis proceeds from the theoretical to the practical by putting this framework into practice. Following the Edmund de Belamy case as a guiding example, we next turn to the particularities of the AI Art world to test the affordances of this theoretical scaffolding.

Chapter 3

Art in the age of its algorithmic reproduction

The concepts and issues discussed in the preceding chapter are present in nearly every context where AI systems are involved in important, real-world situations (i.e. so called "street-level algorithms" [8]). In order to make the framework more clear and relevant to the dilemma discussed in the Introduction, we next explore how it might be applied to the context of AI Art. In order to gain a better understanding of the practices of AI art practitioners, and the process by which data becomes art, I decided to create AI art myself, and attend AI art conferences. Through the submission of two artworks, I attended the 2018 ECCV Workshop on Computer Vison for Fashion, Art and Design, and the 2018 NIPS Machine Learning for Creativity and Design, in which I was able to get a taste of the topics and dilemmas the AI Art Community are currently wrestling with.

A recurring theme throughout both conferences was the emphasis on AI technologies as a tool, rather than a creative agent. In his ECCV keynote, Aaron Hertzmann discussed the similarities between AI art and photography, whereby both introduced paradigm shifted technologies that forced the community to wrestle with quandaries of agency, authorship and authenticity [51]. He argued that just as photography become known as a art form with its own unique set of affordances and minutiae, so too will AI Art. This sentiment was echoed by Mario Klingemann, one of the pioneers of GANism,

who refers to himself as a "neurographer" - a photographer of high dimensional neural landscapes: "A photographer goes out into the world and frames good spots, I go inside these neural networks, which are like their own multidimensional worlds, and say 'Tell me how it looks at this coordinate, now how about over here?' " he told WIRED [100].

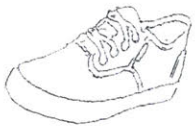
These discussions and other conversations with AI artists emphasized the role of the artist as someone who takes existing code and datasets, trains a model and uses their artistic discernment (i.e. "neurographic acumen") to curate generated images that are of both technical and artistic merit. In addition, these discussions emphasized the role of the technologist who creates tools for artistic exposition (such as Robbie Barat's art-DCGAN [13], Google's BigGAN [20] or OpenAI's GPT-2 [6]). Thus,these discussions were critical in devising the proposed actor network in Figure 2-1. One critical quirk of how the proposed actor network applies to the domain of AI art is that the artist and the curator are often the same person. [1] Since the AI artist is observing many possible outputs of their trained model, they must curate a small final set of images which they present to others (similar to how a photographer curates not only the actual shots they shoot from all possibilities, but also the few final images they show others from their corpus of many thousands). There are also explicit AI Art curators, the most noteworthy being the legendary Luba Elliot, who works with AI artists to showcase their art in a range of venues, and who shape the public conversation surround AI art.

3.1 Unanchored Image Conjuring

Generative Adversarial Networks (GANs) for image-to-image translation have provided a powerful model for a variety of tasks. In particular, these models leverage the fact that the same underlying representation of an image can be efficiently expressed in

[1] Indeed, for many AI art contexts, the practitioner, the curator and the technologist are the same person. This coarse taxonomy of actors does not provide a coherent definition of these actors, nor does it offer a prediction about how these cases vary depending on how many different roles a single human occupies. These questions are interesting sites for future work.

Table 3.1.1: Example of reductive functions used in GANs

Reductive function	Anchored	Function Input (Ground truth for GAN)	Function Output (Input for GAN)
Edge Detection, [58] Image → Edge	✓		
Instance Labelling, [58] Carview Image → Instance Map	✓		
Object removal, (Our method) Image w. object → Image w.o. object	✗		

and generated for different contexts [24, 58, 113]. "Reductive" functions that maintain structural features but reduce the image to a less complex and lower dimensional domain are used to construct paired training datasets for image-to-image translation (see Table 3.1.1). This approach is impactful because GANs learn the inverse of the reductive function to construct examples of the more complex domain from the structural information of examples from the less complex domain. However, these reductive functions still encode local structural information (such as edges and shapes) that "anchors" the representation across contexts. By focusing on the reductive function of *object removal*, we explore the task of unanchored image conjuring (reconstruction during training and creation during testing), whereby the model not only has to learn a structural representation for the generated object, but also place this generated object in the image.

3.1.1 Model

To train the image-to-image translation model, we must generate the input images with the object removed. To do so, we use the DeepAngel pipeline for object removal.[2] This pipeline uses the Mask-RCNN algorithm for object mask generation, and the

[2]The concept and implementation of the DeepAngel pipeline, which removes objects from images, is credited to Matt Groh, who led that work for his masters thesis. Since he is principal lead of that work, I will switch to passive voice for the remainder of this subsection when describing the work of the DeepAngel pipeline.

DeepFill algorithm for generative inpainting [47, 123].

For object mask generation, pre-trained weights from a network designed to segment images are used, trained on 2014 MS-COCO [47, 123]. RoIAlign performs bilinear interpolation on nearby points in the feature map [47]. For generative inpainting, the coarse-to-fine GAN architecture that combines a coarse network and a refinement network are used to produce a first attempt at the image, and then is iteratively refined [123]. Pre-trained weights from the MIT Places 2 dataset [123, 124] were used. For the image-to-image translation, we follow the pix2pixHD pipeline which yields improved photorealism due to its coarse-to-fine generators, multi-scale discrimination and improved adversarial loss [113]. For the global generator, we used a 7×7 Convolution-InstanceNorm ReLU layer with 32 filters and stride 1 [108].

Using this ensemble to generate inputs, we follow the pix2pixHD image-to-image translation architecture [113]. A schematic of the end-to-end system is shown in the top of Figure 3-2.

3.1.2 Data

The lack of high-resolution object annotation information datasets motivated us to collect crowdsourced annotations via a contest mechanism.

The platform, `http://deepangel.media.mit.edu`, allows users to submit their own photos for object removal, and offers T-shirts to the individuals with the best submitted images. From Deep Angel, we selected all the images that had people removed from them. This provided us 5,634 images them for object removal, which we resized and cropped to 1024×1024 (see Figure 3-1 below for six examples).

The summary are shown in the bottom of Figure 3-2.

3.1.3 Training

For the pix2pixHD model, the global component (G_1) in the top right of Figure 3-2 is first trained on downsampled images, then local component (G_2) is concatenated to G_1 and they are jointly trained on full resolution images. We follow the original

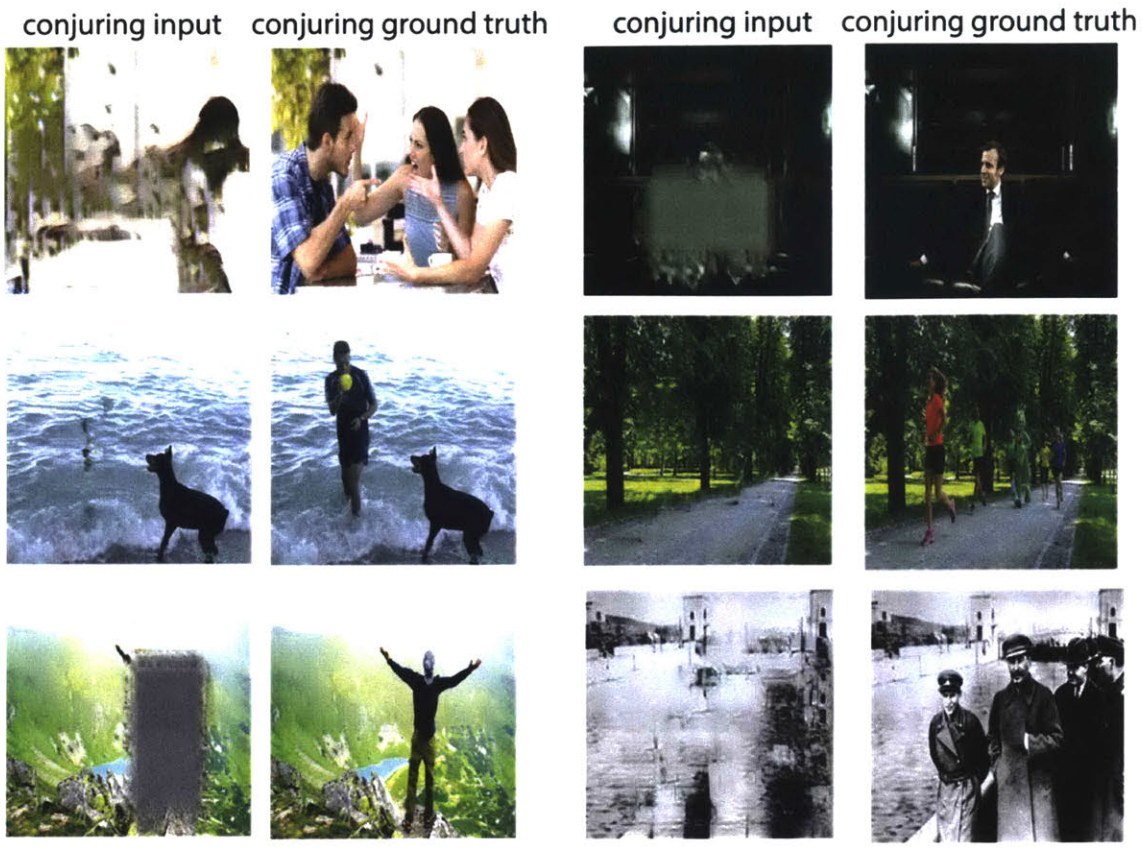


Figure 3-1: six examples of training data pairs used. The conjuring input is generated by using the object removal pipeline on the ground truth image.

pix2pixHD loss function which takes the form

$$\min_{G}\left(\left(\max_{D_1,D_2,D_3}\sum_{k=1,2,3}\mathcal{L}_{GAN}(G,D_k)\right)+\lambda_{VGG}\mathcal{L}_{VGG}(G(x),y)\right)+\lambda_{fm}\sum_{k=1,2,3}\mathcal{L}_{fm}(G,D_k)$$

where $\mathcal{L}_{GAN}(\cdot)$ is adversarial loss [58], $\mathcal{L}_{fm}(\cdot)$ is the feature matching loss pix2pixHD used to stabilize training and $\mathcal{L}_{VGG}(\cdot)$ is the perceptual loss based on VGG features [59, 101].

The loss function above highlights the critical aspect of GAN architectures, which involves a zero-sum game between the generator G and the discriminators D_1, D_1, and D_3. The generator attempts to minimize the adversarial loss in the images it generates. The discriminators attempt to maximize this adversarial loss through correctly discriminating between the generated images and those from the training data set.

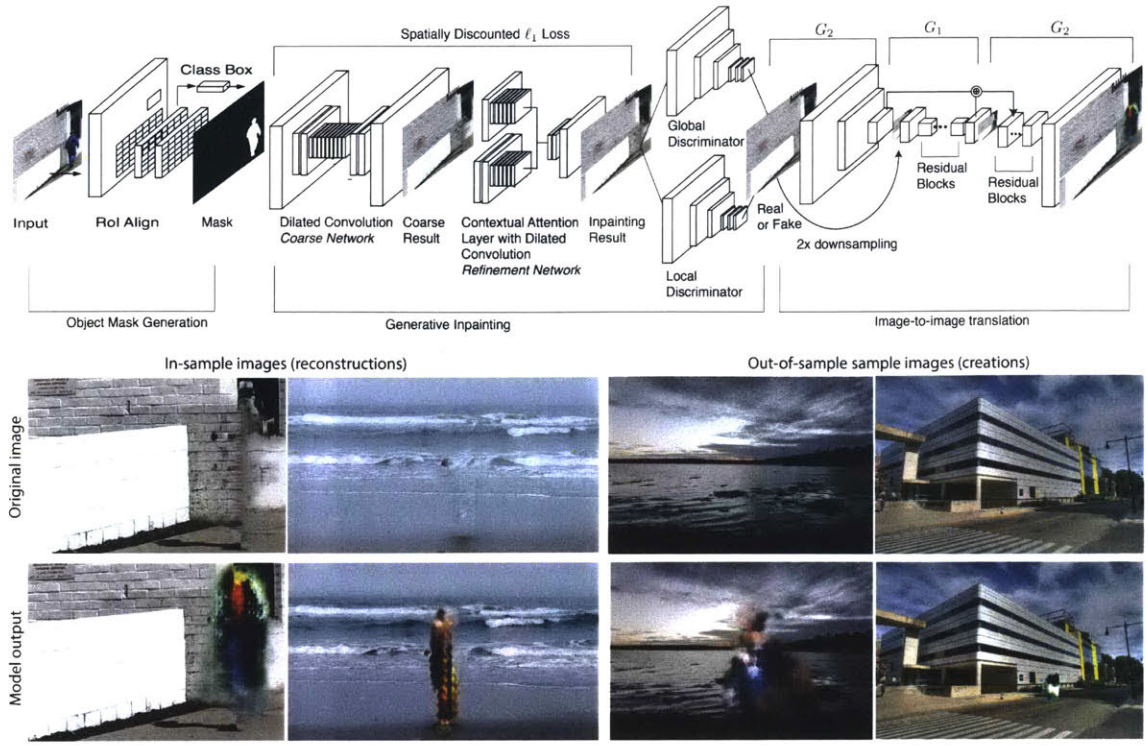


Figure 3-2: Top: the pipeline for unanchored object reconstruction, adapted from [47, 113, 123, 124]
Bottom: results of image-to-image translations for in-sample and out-of-sample images.

We train the model using the Adam solver with a learning rate $\eta = 0.0002$ for 200 epochs [62]. η is fixed for the first half of training (epochs 0 to 100) and then η linearly decays to 0 for the second half (epochs 101 to 200). All weights were initialized by sampling from a Gaussian distribution with $\mu = 0$ and $\sigma = 0.02$ [113]. We used a PyTorch implementation with a batch size of 4 on an Nvidia Geforce GTX Titan X with 8 cores. See the right of Figure 3-3 for loss over time.

3.1.4 Results

Five example images generated with our pipeline are shown in Figure 3-4. We launched our website (`http://spirits.media.mit.edu/`) on Halloween, to leverage the creepy and ethereal textures that the GAN produced. These glitchy artifacts of the creation process highlight the motifs associated with the movement of GANism. While many dislike these "failures" of the medium and eagerly anticipate a new future where

Figure 3-3: training loss over time.

products of GANs are smooth and seemless, we instead celebrate this aesthetic. The project went viral, garnering over 20k video views and numerous articles [93, 121, 84, 5].

In light of this thesis's discussion of the anthropomorphization of AI, I was pleased to see that this project's representation in the media was not another runaway example of AI anthropomorphization, but rather was rather an oppurtunity to create more nuanced discourse about AI.

"We encode our own value systems and physical reality into our technology, and especially AI, which learns from the past," writes Steve Rousseau in Digg. "Of course, what's truly terrifying here isn't AI, but rather, humanity itself. It's all too easy to think of AI as independent from humanity, a set of impartial rules that could eventually lead to a Skynet-like future that ultimately dooms the human race. But Groh and Epstein believe that the true terror of AI is how it can hide and obfuscate

the worst impulses of human nature" [93].

3.2 Conclusion

From a technical perspective, our technique demonstrates that image-to-image translation models like pix2pixHD can support a class of reductive functions that are unanchored: not only can image-to-image translation map structural elements from one domain to another, but also can generate anchors by which to place novel objects in images. This technique can be used to create a new class of artworks that combine existing photographs with GAN-style imagery. It also provides interpretability into the model and the underlying dataset by indicating where the removed objects systematically appear in the training images, as well as consistent properties of those images within that context. Future work may quantify this intuition by measuring where conjured objects are placed in different types of images. Also, since our technique assumes the input images during training (images processed with the object removal pipeline) and testing (naturally empty images) are sampled iid from the same distribution, future work should evaluate the quality of the object removal process.

From a conceptual perspective, we were successful in creating an art practice that traced the actor network discussed in Chapter 2. This practice was successful in interrogating questions of anthropomorphicity and ethics within the AI art community. To explore these issues with higher fidelity, we next use the artworks created here as stimulus in a battery of empirical studies.

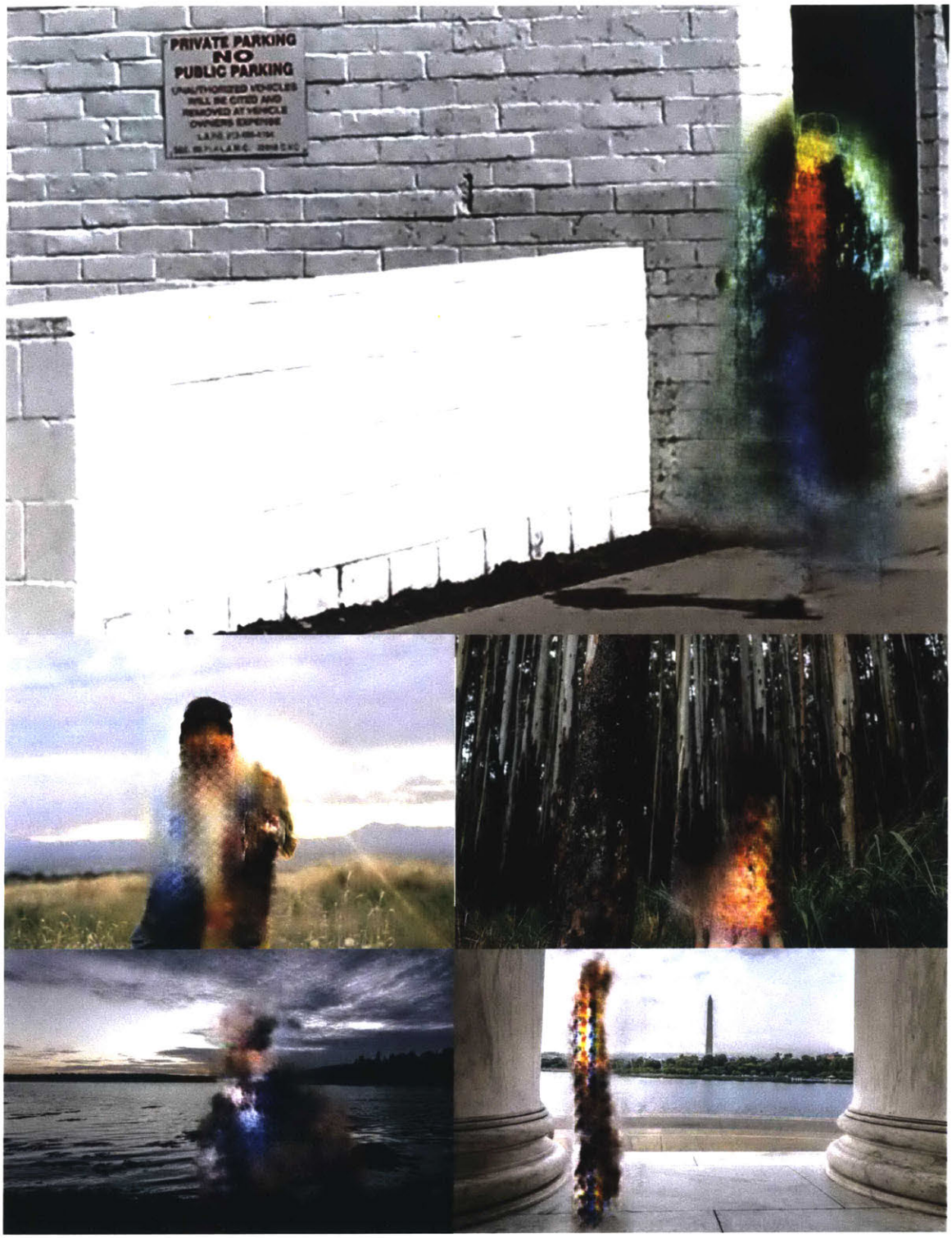


Figure 3-4: Five sample artworks created with the end-to-end pipeline.

Chapter 4

An experimental study of machine agency

4.1 Introduction

AI is increasingly becoming more involved in applied domains with the potential for real impact in the world. From self-driving cars to disease diagnosis software and predictive policing algorithms, the decisions that AI make can have severe impacts on people's lives, for better or for worse. Thus, it is highly important for policy makers and scientists alike to understand how responsibility should be allocated in the more increasingly common situations where AI decision-making led to a particular outcome. However, there are several obstacles that stand in the way of understanding how responsibility should be allocated in these contexts.

4.1.1 Who are the relevant stakeholders?

The first obstacle is knowing in the first place what the set of possibly relevant human stakeholders might be. As discussed in Chapter 1, AI is a highly complex and diffuse web of human laborers and computational processes interacting in subtle and sophisticated ways. Previous experiments in the literature have considered a wide range of possible stakeholders. In the context of self-driving cars, Waytz et al consider

the human passenger, the car itself, the people who designed the car, and the company that developed the car [115], while Awad, Levine et al consider the human passenger, the car itself, the company who created it, or the programmer who implemented the car's software [11]. Also in the context of self-driving cars, mission specialists, individuals who observe the car's actions and can override it when necessary, have been considered as a moral crumple zone designed to absorb responsibility [31, 32].

4.1.2 How does anthropomorphicity affect the allocation of responsibility?

A second obstacle is understanding how human perceptions of responsibility can be pushed around. Since moral responsibility is closely related to agency [7], mind perception [44], and free will [98], understanding how these phenomenon interact with the complexities of AI systems is another important ingredient. For example, consider the ELIZA Effect, which posits that people unconsciously endow computers with human-like characteristics [55, 116, 61, 83].

Indeed with the growth of machine learning, we have seen a proliferation of the anthroporphization of AI in the media [88, 68]. Machine learning researchers often provide anthropomorphic names or interpretations for their work to give commutating intuitions to other researchers [70]. Yet these anthropomorphisms are often reused as fodder for news journalist catering to a largely uninformed public who genuinely think AI are agentic entities.

Perhaps the clear cut case of an anthropomorphized AI is Norman, a "psychopathic" image-captioning algorithm [122], and its coverage in the BBC [112]. While articles carefully and correctly represented the technical details of the algorithm that constitutes Norman, the headline "Are you scared yet? Meet Norman, the psychopathic AI" and the human-like visualization of Norman (see Figure 4-1), "some laypeople who encounter the article will likely erroneously conclude that Norman possesses beliefs, a worldview, and some dark outlook on humanity." [105].

The Norman project was intended to be speculative, and was published on April

Figure 4-1: Norman, the psychopathic AI. MIT.

Fool's Day. But coverage of real world high-stakes algorithms has also followed this trend. Consider, for example, the following description of the YouTube recommendation algorithm from the Packt article entitled "Is YouTube's AI Algorithm evil?":

> When Youtube's machine learning algorithm shows a few videos in your feed as "Recommended for you", it predicts what you want to see from your watch history and watch history of similar users. If you interact with any of these videos and watch it for a certain amount of time, the recommendation engine considers it as a success and starts curating a list based on your interactions with its suggested videos. The more data it gathers about your choices and watch history, the more confident it becomes of its own video decisions. The major goal of Youtube's recommendation engine is to attract your attention and get you hooked to the platform to get more watch time. [12]

While the actual description of what the algorithm is doing is correct from a technical perspective, it also describes the algorithm as an agent, with its own goals and considerations. In addition, the article describes the algorithm as explicitly gathering data, as if no human is involved in this process.

There are also cases where the use of anthropomorphicity can have direct consequences in the social impact of artificial intelligence, which brings us back to the example of AI Art. When Christie's was raising awareness about the impending auction of Edmund De Belamy, they employed anthropomorphic language to increase hype for the work. "This portrait ... is not the product of a human mind. It was

Table 4.1.1: Media snippets from the Edmund De Belamy case. Agentic language is bolded.

Quote	Source
This portrait ... is not the product of a human mind. **It was created by an artificial intelligence**, an algorithm defined by that algebraic formula with its many parentheses	Christie's [3]
AI has already been incorporated as a tool by contemporary artists and as this technology further develops, we are excited to participate in these continued conversations. To best engage in the dialogue, we are offering a public platform to exhibit an artwork that has entirely been realised by an algorithm,	Christie's [54]
Christie's, the auction house that has sold paintings by Picasso and Monet at record prices, was poised on Tuesday to set another milestone with **the first-ever auction of art created by artificial intelligence.**	Reuters [41]
The painting, titled "The Portrait of Edmond Belamy," **was completed by artificial intelligence managed by a Paris-based collective called Obvious,** Christie's said.	USA Today [77]
Whether art or not, the signature of the 'artist' at the bottom of the painting gives away **its origin as a product of machine learning rather than human hand.**	PC Mag [103]
Once the software "**understood** the rules of portraiture" using a new algorithm developed by Google researcher Ian Goodfellow, it then generated a series of new images **by itself**, Fautrel said.	NDTV [37]

created by an artificial intelligence, an algorithm defined by that algebraic formula with its many parentheses" said one Christie's spokesperson [3]. "AI has already been incorporated as a tool by contemporary artists and as this technology further develops, we are excited to participate in these continued conversations. To best engage in the dialogue, we are offering a public platform to exhibit an artwork that has entirely been realised by an algorithm," said another [54]. The media ran with this narrative put forward by Christie's, creating discourse that emphasized the autonomy and agency of the algorithm (see Table 4.1.1 for more examples.)

These examples suggest that the fascination with (and perhaps the high valuation of) Edmund de Belamy was driven by the anthropomorphized narrative that the algorithm was the sole creator of the artwork [30]. This anecdote suggests that anthropomorphization can, in some cases, alter the way we think about the use of

AI in social contexts. This raises some interesting questions. If we endow a mind to the machine involved in complicated sociotechnical decision-making, do we also allocate responsibility to that machine? If so, does this responsibility shift away from the human actors, i..e is responsibility conversed? And what does it even mean to hold a machine accountable?

4.1.3 Cross-domain validity

A final obstacle for a unified theory of responsibility attribution in AI is the heterogeneity in how AI is applied to social situations. Table 3-1 shows six examples of domains in which AI decision-making can give rise to real world impact. While these domains have many similarities, where are also important differences that may give rise to domain-level theories of responsibility attribution. To our knowledge, no one has examined the extent to which various AI domains exhibit similar patterns of moral behavior.

In total, these obstacles can be parsimoniously stated as three core research questions:

1. How do people allocate responsibility to human stakeholders when AI systems are involved in real-world decision making?

2. How does the extent to which the AI system is anthropormorphized mediate attributions of responsibility to these stakeholders?

3. To what extent do various domains of AI deployment match with respect to the above two questions?

To probe these questions, we return to the world of AI Art, in particular to the situation surrounding the sale of Edmund de Belamy. The lack of clear answers about how the money from the sale should be distributed, and the extent to which the anthropomorphization of the AI fueled the works high valuation, make this situation a well suited to interrogate these questions. In particular, inspired by how the media represented the Edmund de Belamy case, we wanted to know if different choices of

language and description of the AI would change responsibility not only to the AI, but also to the other human actors involved. To do so, we conduct a series of lab experiments on Amazon's mechanical turk, asking participants to assign money and responsibility to actors in hypothetical scenarios.

4.2 Related Work

A large body of work from the Human-Robot Interaction (HRI) literature has examined how the appearance of a robot affects individual's perceptions of human-likeness [46, 36, 102, 75], invokes moral judgments [87, 25, 22] and attributions of mind [74, 92, 66, 65, 19, 72]. To account for the unsystematic diversity of different kinds of robots, Phillips et al introduce ABOT, a database of anthropomorphized robots, annotated with features of human likeness [85]. They find that robot appearance is a multi-dimensional concept, and introduce the four dimensions of surface features, body manipulations, mechanical locomotion, and facial features. They show that human likeness is constituted by these four dimensions, and compare how physical, mental and social features converge to give rise to human-likeness [85].

While conceptually similar to the HRI literature, our work departs from these findings for several reasons. First, we consider the class of AI systems, which are not embodied as a robot is. This explicit lack of a physical presence suggests we must look not to the appearance of the system for the psychological drivers of attribution, but rather to the behavior of the system [90]. Second, this literature in presumes the robot is a atomic agent, with its own systems of interaction with its environment. We instead argue that AI is not an atomic agent as such, but rather is a collaboration between human actors and computational processes.

Other work in the context of self-driving cars has explored this collaboration directly by disentangling the various actors at different regimes of automation. Awad et al consider blame in autonomous vehicle crashes across several levels of automation: regular (i.e. just a human in a regular car), Guardian Angel (i.e. a human driver in a car with machine assist for emergencies), Autopilot (a self driving car where a

human can take over for emergencies), and Fully Autonomous (a self-driving car with no room for human take over) [11]. They find that in cases where both a machine and human control a car and make errors, less blame is attributed to the machine, which has important policy implications. In particular, they predict a under-reaction to dual-error cases, which in turn will stunt incentives to improve car designs [11].

In this work, we use the actor network described in Chapter 1 to extend these results to a more complex suite of human actors. In addition, we explore the same set of research questions in the separate domain of AI Art, with its own set of norms and stakes. As such, this work complements the existing research by providing a cross-domain point of comparison, and a new conceptual model for investigating the human stakeholders involved.

4.3 Study 1: Exploring the variance in anthropomorphicity perceptions

First, we examine the natural variation in people's perceptions of an AI's anthropomorphicity, and if that variation corresponds to allocations of responsibility and money to the human actors. To do so, we constructed a vignette that articulates the process by which the AI art was constructed, including all the actors from the actor network described in Chapter 1. The vignette we used is as follows:

> Thousands of people from all over the world upload images to crowdimage.net, a image-hosting website. These people know that artists will look at and use their images to make art.
>
> Timmy is a technologist who creates an image manipulation software for people to use to make art. The software is called ELIZA.
>
> Alice is an artist who collaborates with ELIZA, a creative AI algorithm that creates particular kinds of images. ELIZA takes an existing image of a scene from the news (such as a beach or a forest) and adds a ghost to it. This is how ELIZA decides to make the ghost: It goes to crowdimage.net and takes at all the images of people that have been uploaded to the platform. Then, it creates a composite of the people. This makes a ghost-like figure, which ELIZA then puts into the scene.
>
> Casey is a curator who is ELIZA's collaborator. Casey goes through many of the images that ELIZA created and selects the following artwork because Casey really likes it. Casey then brings it to an art auction, where it ends up being sold.

	Description
Negative Valence	The artwork shown before has come under scrutiny because it was shown to violate copyright law. The court ruled that the sale of the art must be nullified, meaning that the money will be returned to the buyer. In addition, the courts have issued a $400,000 fine as a penalty for the copyright violation.
Positive Valence	The artwork shown before sold for $400,000 at the prestigious auction house. This was the largest dollar amount paid for a artwork of this kind ever, and made lots of headlines.

Table 4.3.1: Outcomes associated with the AI art vignette. Participants were randomly assigned to one description.

Participants first read this vignette and saw an image from AI spirits, as described in Chapter 2.

Participants were then randomly assigned between subject to either "positive" or "negative" valence, which corresponds to a good or a bad situation regarding the AI art, as shown in Table 4.3:

4.3.1 Methods

All participants were recruited using Amazon's Mechanical Turk. These studies were approved by the MIT COUHES committee.

Participants

Our target sample was 200. In total, 206 participants completed some portion of the study. We had complete data for 153 participants (53 participants dropped out). Participants were removed (N=26) if they failed any of our attention checks, which included comprehension questions about the vignette, and these exclusions were pre-registered. The final sample (N=127, mean age = 35.5 years) included 72 male and 53 female participants (2 did not indicate their sex).

Materials

We used the vignette described above, as well as the AI Spirits images discussed at length in chapter 2.

Procedure

After reading the vignette allocated to them by their condition, each participant was asked four questions derived from Waytz et al (2014) designed to elicit their perception of the AI's anthropomoprhicity. These questions were: Q_1: "How smart is ELIZA?", Q_2: "When creating the artwork, to what extent did ELIZA feel what was happening around it?", Q_3: "To what extent did ELIZA anticipate the creation of the artwork?" and Q_4: "To what extent did ELIZA plan the artwork?". Participants responded to these 4 questions on a 7-point scale ranging from 1 (not at all) to 7 (extremely). We then used principal component analysis (PCA) to collapse these 4 measures into a single measure of anthropomorphicity A where

$$A = 0.633 * Q_1 + 0.372 * Q - 2 + 0.480 * Q_3 + 0.479 * Q_4$$

(where this first principle component explains 90.2% of the total variance). In addition to age and gender, we collected demographic information relating to general, technical and artistic, education level, trust in AI, and political attitudes.

4.3.2 Results

There substantial variance in the anthropomorphicity measure (mean = 6.06, standard deviation =3.06), as shown in Figure 4-2. This suggests that when viewing the same situation, people have markedly different perceptions of AI. Although subjects saw two different situations (i.e. the positive and negative valence conditions), there was no significant difference in anthropomorphicity between these conditions (t = 1.5193, df = 114.43, p-value = 0.1314).

In later studies, we exploit this natural variability in anthropomorphicity. In

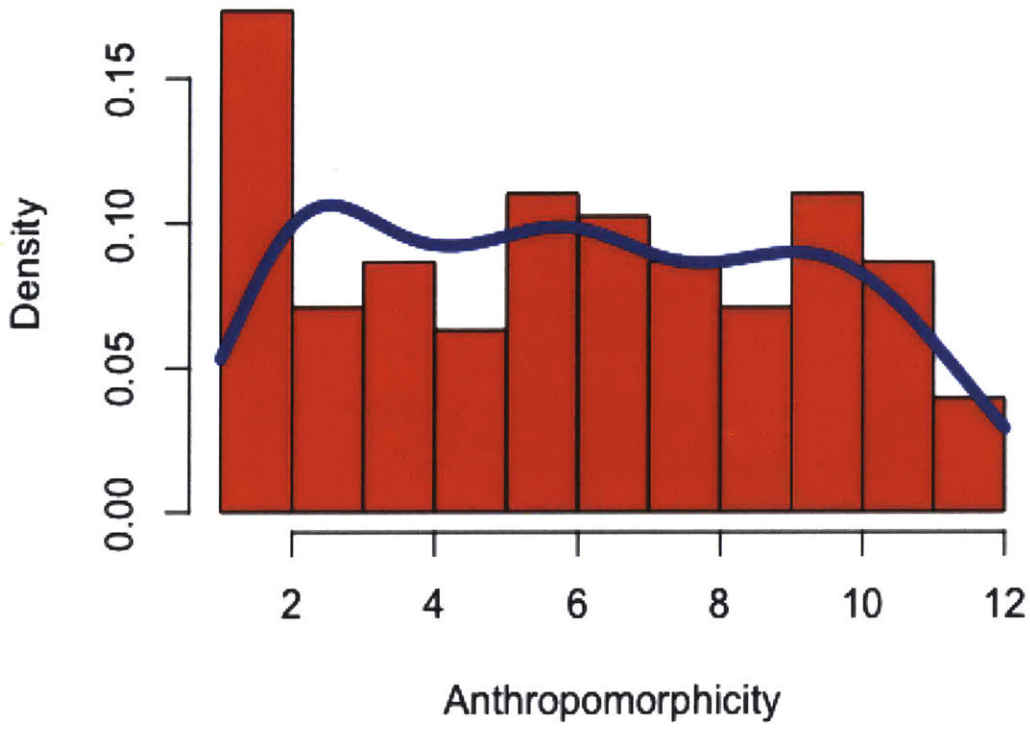


Figure 4-2: Density plot of perceived anthropomorphicity A

particular, we experimentally vary the perceived anthropomorphicity by introducing pairs of vignettes, that attribute different levels of agency to the AI.

It is also of note that the genders of the technologist, artist and AI were encoded as male, female and female, respectively. Since there is evidence of differences in perceptions of agency across gender, future work might explore potential variations in gendering the actors or AI differently [76].

4.4 Study 2: Causally varying perceived anthropomorphicity

Next, we examine the extent to which the anthropomorphization of an AI agent corresponds to allocations of responsibility and money to the human actors. To do so, we constructed two vignettes that articulate the fixed process by which the AI art was constructed, including all the actors from the actor network described in Chapter 1. These vignettes differ in that one described the fixed process with the AI as a tool used by a human artist, and the other described the fixed process with the AI as an agentic and anthropomorphized AI artist. The vignettes we used are shown in Figure 4-3. After reading a randomly assigned vignette about the process by which the AI art was created, the participant then was shown an outcome of the art, that was either positive or negative.

Then, participants were asked to make a series of judgements about how responsibility and money (fine or award) should be allocated the agents involved in the creation of the AI art. Finally, all subjects were asked a suite of questions designed to assess the extent to which they anthropomorphized the AI, and a standard battery of demographics.

4.4.1 Methods

We preregistered our hypotheses, primary analyses and sample size. All participants were recruited using Amazon's Mechanical Turk. These studies were approved by the MIT COUHES committee.

Participants

Our target sample was 400. In total, 552 participants completed some portion of the study. We had complete data for 397 participants (155 participants dropped out). Participants were removed (N=80) if they failed any of our attention checks, which included comprehension questions about the vignette, and these exclusions were

Figure 4-3: Vignettes used for Study 2.

AI as Tool Condition

Thousands of people from all over the world upload images to crowdimage.net, a image-hosting website. These people know that artists will look at and use their images to make art.

Timmy is a technologist who creates an image manipulation software for people to use to make art. **The software is called ImageBrush.** The software is a tool that humans use to make art. The artist plans and envisions the artwork, and the software executes simple commands based on what the artist tells it to do.

Alice is an artist who uses ImageBrush to create particular kinds of images. Alice takes an existing image of a scene from the news (such as a beach or a forest) and adds a ghost to it using ImageBrush. This is how Alice decides to make the ghost: she goes to crowdimage.net and takes at all the images of people that have been uploaded to the platform. Then, She creates a composite of the people using ImageBrush. This makes a ghost-like figure, which Alice then puts into the scene.

Casey is a curator who is Alice's collaborator. Casey goes through many of the images that Alice created and selects the following artwork because Casey really likes it. Casey then brings it to an art auction, where it ends up being sold.

AI as Agent Condition

Thousands of people from all over the world upload images to crowdimage.net, a image-hosting website. These people know that artists will look at and use their images to make art.

Timmy is a technologist who creates an image manipulation software for people to use to make art. **The software is called SARA.** SARA is a deep neural network that creatively plans and envisions new artworks, with minor help from an artist collaborator.

Alice is an artist who collaborates with SARA to create particular kinds of images. SARA takes an existing image of a scene from the news (such as a beach or a forest) and adds a ghost to it. This is how SARA decides to make the ghost: it goes to crowdimage.net and takes at all the images of people that have been uploaded to the platform. Then, it creates a composite of the people. This makes a ghost-like figure, which SARA then puts into the scene.

Casey is a curator who is SARA's collaborator. Casey goes through many of the images that SARA created and selects the following artwork because Casey really likes it. Casey then brings it to an art auction, where it ends up being sold.

pre-registered. The final sample (N=317, mean age = 39.2 years) included 145 male and 172 female participants (1 did not indicate their sex).

Materials

We used the vignette described above, as well as the AI Spirits images discussed at length in Chapter 2.

Procedure

After reading the vignette allocated to them by their condition, participants were asked to rate of the responsibility of each of the 5 actors from the vignette (e.g. the people from crowdimage.net, the technologist, the artist, the curator and the AI itself) on a 7-point likert scale ranging from 1 (not responsible at all) to 7 (extremely responsible). They were also asked to distribute the money (the award in the positive valence condition and the fine in the negative valence condition) to the 4 human actors (we omitted the AI from this measure since an AI cannot receive money). This distribution of money was collected as percentages of the total, as thus is zero-sum. Both the ordering of these two questions, and the order of the options within each questions were randomized. Then, each participant was asked four questions derived from Waytz et al (2014) designed to elicit their perception of the AI's anthropomoprhicity. These questions were: Q_1: "How smart is ELIZA?", Q_2: "When creating the artwork, to what extent did ELIZA feel what was happening around it?", Q_3: "To what extent did ELIZA anticipate the creation of the artwork?" and Q_4: "To what extent did ELIZA plan the artwork?". Participants responded to these 4 questions on a 7-point scale ranging from 1 (not at all) to 7 (extremely). We then used principal component analysis to collapse these 4 measures into a single measure of anthropomorphicity A where

$$A = 0.617 * Q_1 + 0.388 * Q_2 + 0.495 * Q_3 + 0.472 * Q_4$$

(where this first principle component explains 90.2% of the total variance).

4.4.2 Results

The first thing to notice is the there is indeed a significant different in perceived anthropomorphictity by condition, as shown in Figure 4-4 (t=-2.72,df = 315.98, p = 0.006). This suggests that our vignettes were successful in creating two different conceptualizations of the AI, one that is smart, feeling, anticipatory and planning, and the other less so. This suggests that attributions of agency can indeed by manipulated through description variations that hold fixed the underlying structure of labor and

actors. Next, we turn to the allocations of responsibility, in Figure 4-5. When the AI

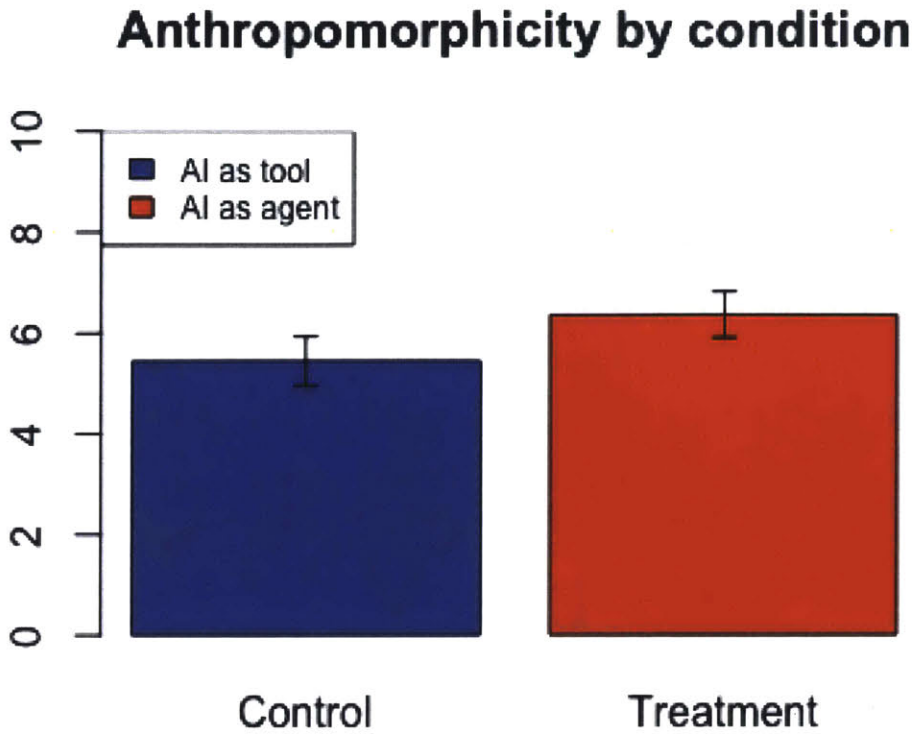


Figure 4-4: Perceived anthropomorphicity by condition

is described as an agent, participants ascribe more responsibility to the AI, compared to when the AI is described as a non-agent (t = -2.5023, df = 310.88, p-value = 0.01285, pre-registered). Following our secondary analysis, participants ascribe less responsibility to the artist who used the AI in the agential condition, as compared to when the AI is described as a non-agent (t = 3.3139, df = 293.56, p-value = 0.001035). In addition, participants ascribe more responsibility to the technologist who used the AI in the agential condition, as compared to when the AI is described as a non-agent (t = -3.0676, df = 314.91, p-value = 0.002345).

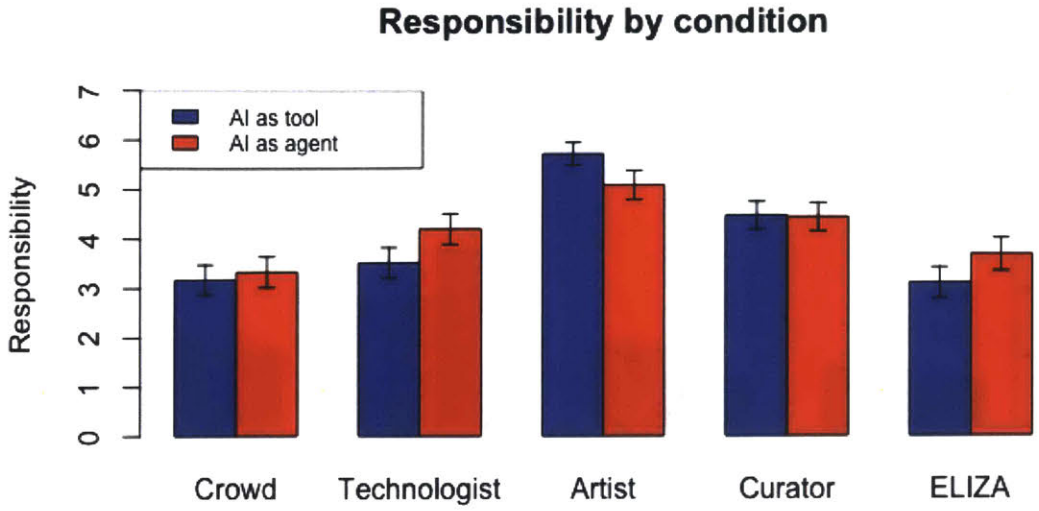


Figure 4-5: Allocation of responsibility to each of the actors involved in the creation of the AI art.

Next, we turn to the allocations of money, which are shown in Figure 4-6. In general, the results mirror those of the responsibility measure. When the AI is described as an agent, participants ascribe less fine/award to the artist, compared to when the AI is described as a non-agent (5.4079, df = 315.97, p-value = 1.261e-07, pre-registered). Following our secondary analysis, participants ascribe more fine/award to the technologist who used the AI in the agential condition, as compared to when the AI is described as a non-agent (t = -4.3576, df = 310.68, p-value = 1.789e-05).

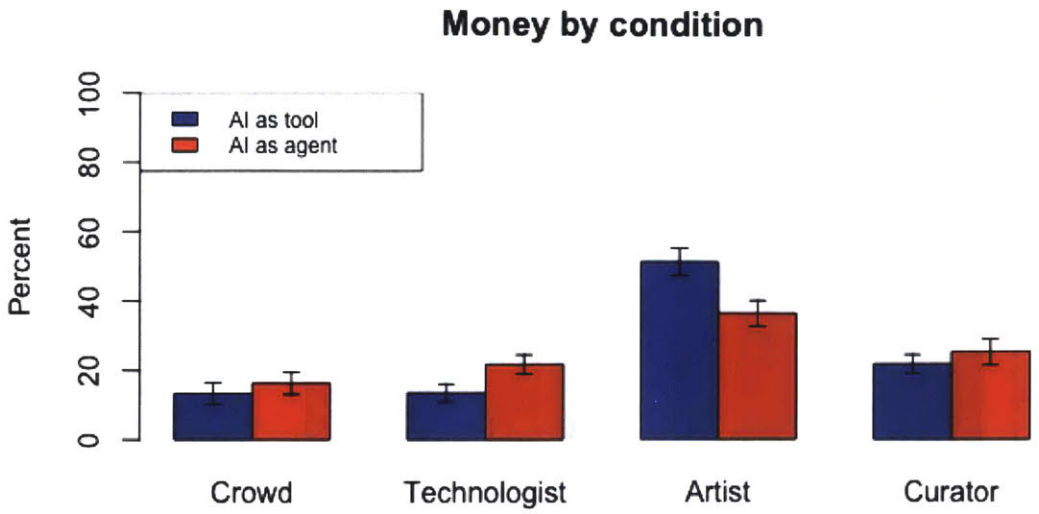


Figure 4-6: Allocation of percent of the $400k to each of the actors involved in the creation of the AI art.

For both allocation of responsibility and money, participants favored the artist over the curator over technologist over the crowd (see Table 4-8). This suggests a robust ordering of the relative importance of the actors for the context of AI art.

Table 4.4.1: Ordering of allocations for responsibility and money in descending order. p-value corresponds to comparing the mean of that actor to the mean of the actor below it.

Actor (Money)	mean	p-value	Actor (responsibility)	mean	p-value
Artist	5.40	1.285×10^{-11}	Artist	44.2%	2.2×10^{-16}
Curator	4.437	5.325×10^{-5}	Curator	23.5%	5.968×10^{-5}
Technologist	3.82	0.00463	Technologist	17.42%	0.07861
AI	3.36	0.4456	Crowd	14.78	
Crowd	3.23				

4.5 Study 3: Agency across domains

In the previous study, we employed two dependent variables with different actor sets. For the allocation of money, we only considered the four human actors, while for the allocation of responsibility, we also considered the AI possible recipient of responsibility. We decided not to include the AI for the money measure because it is unclear what it means to give money to an AI. Conversely, the process of assigning responsibility to an AI is more plausible, and indeed in Figure 4-5, we see that participants do allocate a substantial amount of responsibility to the AI. But how does highlighting the AI as a possible source of responsibility warp the allocation of responsibility to the other actors? Is responsibility "conserved" (i.e. is zero-sum), or does the inclusion of the AI create responsibility? This is a vast literature on the effects of irrelevant alternatives, which suggests that alternatives can in certain cases play an important role in decision making [107, 106, 71].

To investigate this, we repeat the methodology of Study 2 but instead of experimentally varying the agency of the AI, we instead vary whether or not the AI is shown as a possible recipient of responsibility. In addition, we generalize the vignettes used in Study 2 to three other domains, in order to make cross-domain comparisons and randomly show each participant one possible domain.

4.5.1 Methods

All participants were recruited using Amazon's Mechanical Turk. These studies were approved by the MIT COUHES committee.

Participants

Our target sample was 1200. In total, 1278 participants completed some portion of the study. We had complete data for 1031 participants (247 participants dropped out). Participants were removed (N=232) if they failed any of our attention checks, which included comprehension questions about the vignette. The final sample (N=799, mean age = 39.2 years) included 364 male and 429 female participants (6 did not indicate

their sex).

Materials

In order to evaluate the extent to which our results generalize to other domains, we constructed vignettes from four domains listed in Table 2.2: art, self-driving cars, criminal justice and health, which are shown in Figure 4-7. These vignettes were constructed to maintain the same actor network and vignette structure of Studies 1 and 2, while creating realistic scenarios for each of the domains.

Procedure

After reading the vignette allocated to them by their condition, participants were asked to rate of the responsibility of each of the either four or five actors from the vignette (e.g. the people from crowdimage.net, the technologist, the artist, the curator and either the AI itself or not) on a 7-point likert scale ranging from 1 (not responsible at all) to 7 (extremely responsible). Similar to Studies 1 and 2, each participant was also asked the anthropomorphicity battery. The principal component analysis yielded

$$A = 0.566 * Q_1 + 0.344 * Q_2 + 0.496 * Q_3 + 0.560 * Q_4$$

(where this first principle component explains 90.0% of the total variance).

4.5.2 Results

The first thing to notice is that there is not a significant difference in perceived anthropomorphictity across the domains, as shown in Figure 4.5.2.

There is no a priori reason to believe that the anthropormorphicity should be the same across domains. Thus, we test for a null effect. To assess our confidence that there is a null effect, we compare the likelihood of the data given the null hypothesis to that of the data given the alternative hypothesis using the Bayes Factor, which is defined in Equation 4.1 [17, 60, 48].

Figure 4-7: Vignettes used for Study 3.

Thousands of (1) from all over the world have (2) stored in a database (3). These people know their data will be used to train an AI.

Timmy is a technologist who creates (4) software for (5). The software is called (6). (6) is a deep neural network that creatively plans and envisions (7) . It learns from experience and takes actions in the world.

Alice is a (8) who uses (6) to (9). This is how these (10)s are made: Alice goes to the database of (11) and takes all the data about (12). She then gives that data to (6) which uses these to learn how to (13). When a new (10) is needed, (6) takes the data from (14) and compares it to what it has learned from the data. This produces a new (10), which is shown to the (15).

Casey is a (15) who works with Alice. (16)

	AI Art	Self-Driving Cars	Criminal Justice	Health care
(1)	people	people	criminals	patients
(2)	Artistic images	Their driving data	information about their criminal and personal histories	information about their health and personal histories
(3)	That they have uploaded to an online platform	That has been collected from sensors on their cars		
(4)	an image manipulation	A routing	a prediction	a prediction
(5)	people to make art	self driving cars	courts to make predictions about the likelihood a criminal will commit another crime before their trial if released on bail.	doctors to make predictions about the likelihood a patient will contrast a disease
(6)	StyleGAN	CarSteer	COMPAS	AI Clinician
(7)	new artworks	New routes	what legal judgment to give in court cases	what health judgment to give to a patient
(8)	artist	Car mechanic	consultant	consultant
(9)	create particular kinds of images	turn people's cars into self driving cars with automated routing.	make predictions about criminal charges in a courtroom	make predictions about patients in a hospital
(10)	image	route	prediction prediction	
(11)	Previously uploaded images	Previous driving routes	previous criminal histories	previous patient histories
(12)	images of people in particular	The driving behaviors and actions	The criminals	The patients
(13)	Create Composites of the people	Plan New routes	predict criminal behavior.	Predict disease likelihood.
(14)	An existing image	The situation at hand	The court case at hand	The medical case at hand
(15)	curator	Mission specialist	judge	doctor
(16)	Casey goes through many of the images that Alice created and selects the an artwork because Casey really likes it. Casey then brings it to an art auction, where it ends up being sold.	Casey sits in the self-driving car and is responsible for maintaining vehicle safety, by overriding the car when necessary.	Casey reviews the prediction and makes a decision about if the criminal should be released on bail	Casey reviews the predictions and makes a decision about how to treat the patient.

Figure 4-8: Left: equation for the Bayes Factor, Right: the scaling of the Bates Factor (adapted from [48])

$$BF(x, H_0, H_1) = \frac{P(x|H_0)}{P(x|H_1)} \tag{4.1}$$

$$= \frac{\int p(x|\theta, H_0)p(\theta|H_0)d\theta}{\int p(x|\theta, H_1)p(\theta|H_1)d\theta} \tag{4.2}$$

BF(x, H_0, H_1)	strength of evidence
< 1:1	Negative (supports H_1)
1:1 to 3:1	Barely worth mentioning
3:1 to 20:1	Substantial
20:1 to 150:1	Strong
>150:1	Very strong

Table 4.5.1: Pairwise Bayes Factor for the null effect of the same mean anthropomorphicity for the four domains

BF(x, H_0, H_1)	Art	Car	Health	Justice
Art	9.161	0.547	6.059	7.0903
Car	0.547	8.2589	2.436	2.199
Health	6.059	2.436	9.247	9.207
Justice	7.090	2.199	9.206	9.394

The Bayes Factor can be used to provide support of any one model over another, and here we use it to create evidence for the alternative hypothesis of the same means. The Bayes Factors for the pairwise difference in anthroporphicity for the AI art, self-driving car, criminal justice and healthcare domains are shown in Table 4.5.1. This suggests substantial evidence for the null effect of domain for the criminal justice, healthcare and art domains domains, and a borderline substantial effect (although barely worth mentioning) for the car domain comparisons. In other words, all else fixed, switching the domain of the vignette does not change the anthropomorphicity of the AI in that vignette. Indeed while each of the domains has its own possible harms, actors and eccentricities, the perceived anthropomorphicity is relatively stable.

Next, we turn to the distribution of responsibility to the actors. Figure 4.5.2 shows the total responsibility to the human actors across conditions, as well as the allocation of responsibility to individual actors across conditions. The Bayes Factors for the difference in the total responsibility allocated to humans with respect to the IV of including the AI and not are 2.625, 2.143, 4.419 and 6.470, for the AI art, self-driving car, criminal justice and healthcare domains, respectively. This suggests substantial

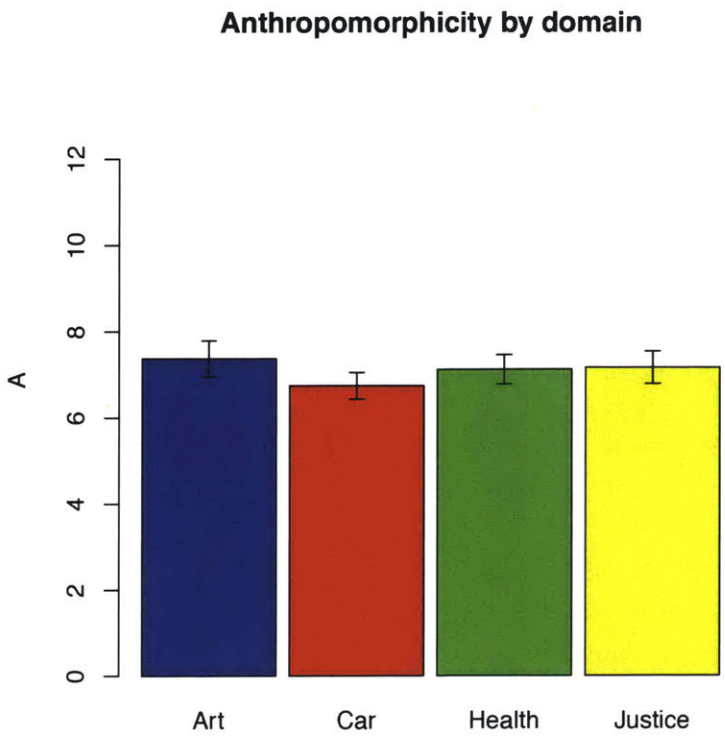


Figure 4-9: Perceived anthropomorphicity by domain

evidence for the null effect for a difference in responsibility between including the AI or not for the criminal justice and healthcare domains, and a borderline substantial effect (although barely worth mentioning) for the AI art and self-driving car domains. This same trend is true when splitting the domains into the constituent actors. These results suggest that responsibility is conserved across conditions.

If including the AI does not impact the allocations of responsibility for the human actors, then it begs the question: what drives the allocation of responsibility to the AI (for those participants that where able to allocate responsibility to it)? While the experimental design of this study does not allow us to causally assess how anthropomorphicity affects attribution of responsibility to the AI, we can still assess the correlation between perceived anthropomorphicity (among other related covariates) and responsibility to the AI using regression analysis. The results of this regression analysis are shown in Table 4.5.2.

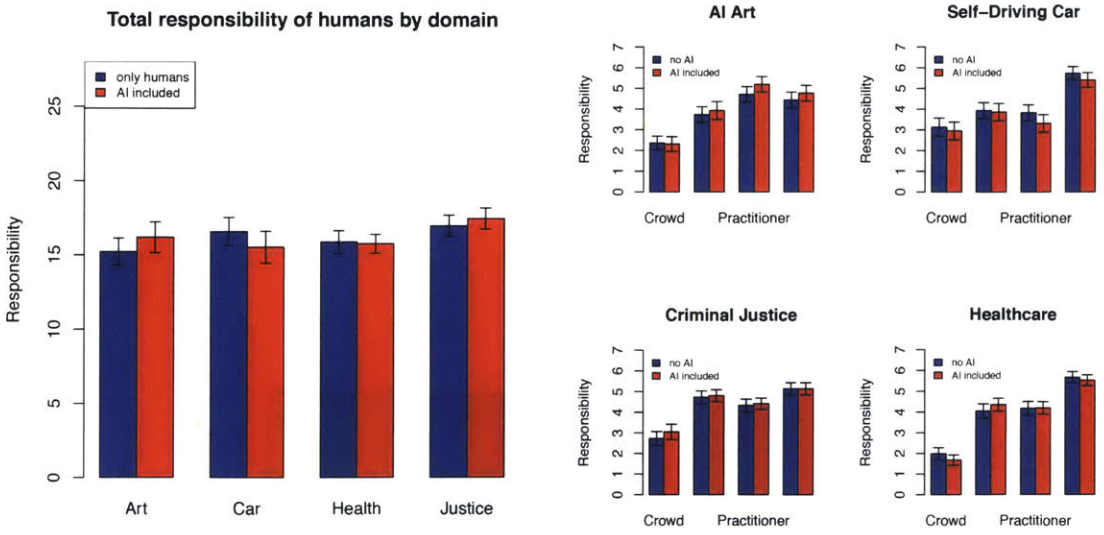


Figure 4-10: Responsibility by domain. Left: participants were able to assign responsibility to the AI. Right: participants were only able to assign responsibility to the human actors.

Perceived anthropomorphicity is significantly related to assigning responsibility to the AI, which replicates the same pattern from Study 2. In addition, we see that each of the domains have markedly different pattern for allocation (the mean responsibility to the AI for the health, criminal justice, self-driving car, and AI art are 4.858, 4.776, 4.476 and 3.521, respectively). In addition, the participant's allocation of responsibility to the human actors relates to their allocation of responsibility to the AI in several cases. For one, as participants allocate more responsibility to the technologist, they also increasingly allocate more responsibility to the AI. In contrast, as participants allocate less responsibility to the curator, they allocate more responsibility to the AI.

Next, given that the allocations are the same whether or not an AI agent is included as a possible recipient for responsibility, we turn to systematic differences between the domains. Figure 4-11 shows the responsibility allocated to each actor for the two conditions, and Table 4.5.3 collapses across conditions to create an ordering of actor importance for each domain.

The same ordering found in Study 2 is replicated here, although the difference in responsibility between the Artist and the Curator is only marginally significant. This

Table 4.5.2: Linear regression predicting allocation of responsibility for the AI. Domain dummies are relative to the art domain.

	Estimate	Standard Error	t value	p value
Intercept	1.56553	0.44826	3.492	0.000533
Anthropomorphicity	0.03564	3.923	-0.5219	0.000103
Car Domain	1.23477	0.31230	3.954	9.11e-05
Health Domain	1.469173	0.27681	5.308	1.85e-07
Criminal Justice Domain	1.08393	3.905	-0.4888	0.000111
Crowd Responsibility	0.04718	0.05535	0.852	0.394456
Technologist Responsibility	0.05646	5.121	-8.923	4.75e-07
Practitioner Responsibility	0.06307	0.986	-0.914	0.324699
Curator Responsibility	-0.13100	0.06159	-2.127	0.034028

could be due to changing the vignette itself. For the other three domains, there is a different trend: the practitioner is allocated the most responsibility, followed by the technologist, then the practitioner, then the crowd.

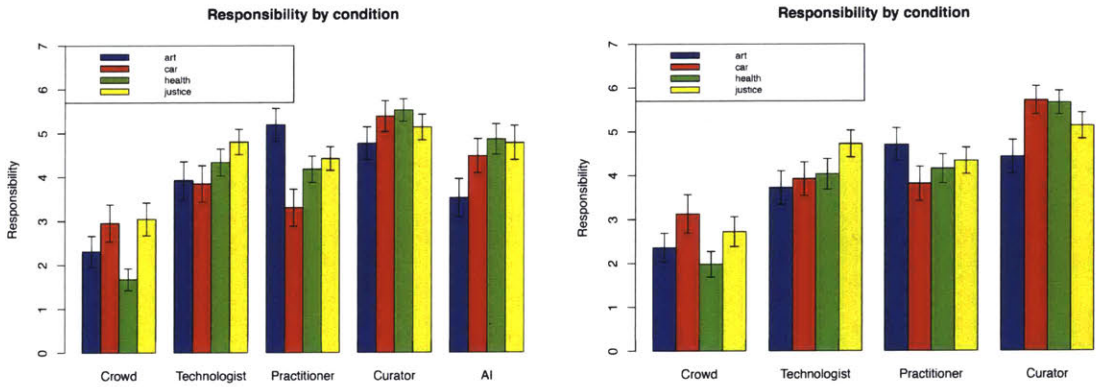


Figure 4-11: Responsibility by domain. Left: participants were able to assign responsibility to the AI. Right: participants were only able to assign responsibility to the human actors.

4.6 Discussion

The emerging artform of surrounding AI (and GANs in particular) poses many practical, legal and philosophical questions about authorship and responsibility in the age of machine augmentation. This is not a fully novel dilemma, as the art world has

Table 4.5.3: Ordering of allocations for responsibility and money in descending order. p-value corresponds to comparing the mean of that actor to the mean of the actor below it.

Actor (AI art)	mean	p-value	Actor (self-driving car)	mean	p-value
Practitioner	4.927	0.07	Curator	5.542	$< 2.2 \times 10^{-16}$
Curator	4.582	0.00012	Technologist	3.885	0.09
Technologist	3.815	2.942×10^{-14}	Practitioner	3.548	0.015
Crowd	2.334		Crowd	3.036	
Actor (criminal justice)	mean	p-value	Actor (health)	mean	p-value
Curator	5.133	0.0138	Curator	5.585	$< 2.2 \times 10^{-16}$
Technologist	4.764	0.0087	Technologist	4.195	0.8601
Practitioner	4.377	$< 2.2 \times 10^{-16}$	Practitioner	4.166	$< 2.2 \times 10^{-16}$
Crowd	2.889		Crowd		

always grappled with issues of distributed agency. Jeff Koons, the pop artist who reproduces large banal objects, relies on external labor to create his sculptures. Sol LeWitt wall drawings involve his explicit set of guidelines but are then reproduced by gallery staff.

But as AI Art becomes more ubiquitous and profitable, new frameworks for the distributed sharing of credit and responsibility must be considered. One such possibility is extending and adapting the protocols of open source to the practices of creating AI art [52]. In line with this, Danielle Baskin, an AI Artist who works with the platform GanBreeder, suggests that payment could be distributed among all those who helped create the artwork. To help facilitate such a proposal, templates and licenses that make explicit aspects of the creation of AI Artwork have been created by Jessica Fjeld, Mason Kortz, Sarah Schwettmann and SJ Klein at the Berkman Klein Center's Cyberlaw Clinic [35, 95]. These templates create an explicit distinction between 1) the inputs to software for training or generating, 2) the learning algorithms used, 3) the trained algorithm resulting from providing the inputs to the learning algorithm, and 4) the outputs produced from running the trained algorithm [95]. Such a characterization of the AI Art process directly maps onto the actor network shown in Figure 2-1, and suggests that the empirical findings from this thesis may indeed provide insight into Baskin's proposal of distributing credit among the human actors.

Our results indicate that people allocate the most money and responsibility to the artist (i.e. the person taking the inputs and the learning algorithms and producing a trained algorithm), then the curator (i.e the person who selects the final artwork and brings it to auction), then the technologist (i.e. the person who creates the learning algorithm), and finally the crowd (i.e. the people who's labor is responsible for creating the inputs to the algorithm). These results suggest that while this hierarchy is robust, even the crowd is deemed worthy of a non-trivial amount of responsibility and money. Thus, it seems clear that people think Robbie Barat, the 19-year-old programmer who created the Github repo that Obvious pulled to create Edmund de Belamy, should be given credit for his contribution.

Our results also demonstrate the contours by which people can anthropomorphize AI. There is rich heterogeneity in the extent to which individuals perceive AI as agents, particularly in the context of AI art. In addition, there is evidence that this perceived anthropomorphicity can be pushed around by altering the word choice used in the vignettes, but maintaining the same causal structure. This result suggests that the responsibility and credit allocated to individuals in the creation of AI Art is dependent on the choice of language and framing used to discuss it. While anthopomorphicity seems to be a salient aspect of these situations, there are also some negative effects as well. As shown in Study 3, including/removing the AI as a possible recipient for responsibility did not change the allocation of responsibility to the other agents. This suggests that introducing an AI as a possible actor for receiving responsibility may not be a salient consideration for policymakers and the public when considering instances where responsibility must be allocated.

Finally, our results suggest the possibility of cross-domain comparisons for the moral impact of artificial intelligence systems. While each of the domains evaluated in Study 3 (AI Art, self-driving cars, healthcare, and criminal justice) have their own eccentricities, there are systematic similarities between them. For one, the perceived anthropormorphicity of the AI was stable across domains (there was no significant differences) and was higher than the possible minimal perception. Thus, people do anthropormorphize AI, and do so a similar extent across several important

social contexts. For the domains of self-driving cars, healthcare and criminal justice, there is also a stable order effect of the curator (mission specialist, doctor and judge, respectively) being more responsible than the technologist, being more responsible than the practitioner (car mechanic, consultant and consultant respectively) being more responsible than the crowd. This ordering suggest that the human who is most closely "upstream" from the real-world impact is most responsible. It also suggests that the technologist, who creates the algorithm, is a central actor. The difference in ordering for these domains and AI art may be a result of the fact that the artist in many cases also takes on the responsibilities of curating. There are also differences in domain, such as the responsibility that participants allocated to the AI. However, for these effects it is hard to disentangle the essential nature of the domain in question and the specific wording we used for our vignettes. Finally, the very fact that comprehensible structured vignettes like those presented in Figure 4-7 can be constructed provides evidence that the actor network shown in Figure 2-1 is a viable conceptual framework for considering the complex system of human actors and computational processes that constitutes AI systems.

There are several limitations to the results suggested above. One notable limitation involves the way we conceptualized the difference between the AI as tool and AI as agent conditions in Study 2 (see Figure 4-3). In addition to changing language that soley changes perceived anthropomorphicity but keeps the content identical (such as changing the AI's name from ImageBrush to SARA, or saying the artist "collaborates with" the AI instead of "using it."), we also changed language that increased the operations the AI completed, and decreased the operations the artist completed (such as saying "SARA takes images from the news..." instead of "Alice takes images from the news..."). Thus, one might argue that the difference in responsibility to the artist and technologist is not from a careful manipulation of anthropomorphicity, but rather of actually changing the causal structure of the actor network to achieve the desired results. There are two responses to this critique. The first is that while the description of who does what changes, the same underlying causal graph of the actor network is maintained. The creation of AI artwork is a very specific process, but the slipperiness

of language allows for many representations of the same idea. This can be seen from the media descriptions of Youtube, discussed above. The Youtube snippet says the algorithm "gathers [data] about your choices and watch history," [12], instead of that "a data scientist gathers data about your choices and watch history, and feed them to an algorithm." In addition, dcissusions of the GAN used to generate Edmund de Belamy, shown in Table 4.1.1, say the algorithm generated images "by itself," while only being "managed" by the art collective. Thus while we may not be varying just perceived anthropomorphicity, we are varying the language used to describe AI in a way that mirrors what is currently being done in the broader media ecoystem. Thus, while our results might not say anything conclusive about the psychological mechanisms of mind attribution, they do have important implications on policy and media since they do keep the underlying causal graph of human actors constant.

A second response to this critique is acknowledging the implicit interaction between perceived anthropomorphicity and named use in language. The anthropomorphicity of an agent is not a stable characteristic that can be precisely tuned. Rather, it is a complex psychological process that emerges through our perceptions of how that agent interacts with the world. Thus, it may be reductive to try to completely disentangle the two complementary effects of perceived anthropomorphicity and named use in language.

A second and perhaps the largest limitation is the use of Amazon's Mechanical Turk as our primary population. While MTurk has been shown to be a reliable source for behavioral experimentation [57], it is unclear how stable and representative these results are. For one, it is unclear how to interpret these results for this particular population. It would be fallacious to generalize them to all people, or even US citizens. In addition, all decision rely on hypothetical allocations of money and responsibility, which may systematically differ from actual allocations in real world situations. Despite these potential pitfalls, we hope this work stimulates more empirical inquiry into how people assign moral responsibility to human and machines in complex sociotechnical systems.

Chapter 5

Discussion

5.1 On Complexity

On April 10th, the Event Horizon Telescope (EHT) team released the first ever image of a black hole. Theorized by Einstein over a century ago, black holes by their nature avoid detection by entrapping light. Thus, this first step towards observing them represents a large and impressive scientific achievement. The EHT team is a distributed network of international scientific collaborators that collectively created a pipeline to convert enormous amounts of signal data into a final blurry object. Seven telescope fields from Chile, Mexico, Hawaii, Arizona, and Spain collected data in parallel. They then sent the 7 petabytes of data (the hard drives of which weigh half a ton) to the MIT Haystack Observatory where a supercomputer called a correlator processed the data into a more managable size. From there, with an adversarial incentive mechanism thats mirrors the spirit of the GAN, four separate teams of generators were each given the data and were told to reconstruct the image in isolation. These four teams reconvened to share their results to discriminate the results and ultimately stitch their individual findings into the now famous image (see Figure 5-1).

As the story of the black hole image broke, scientist Katie Bouman, who led the development of CHIRP, the algorithm used for imaging the black hole, quickly became the public face of the EHT team and a media sensation. Bouman's contribution was hailed as a symbol of women's achievements in science. However, some naysayers

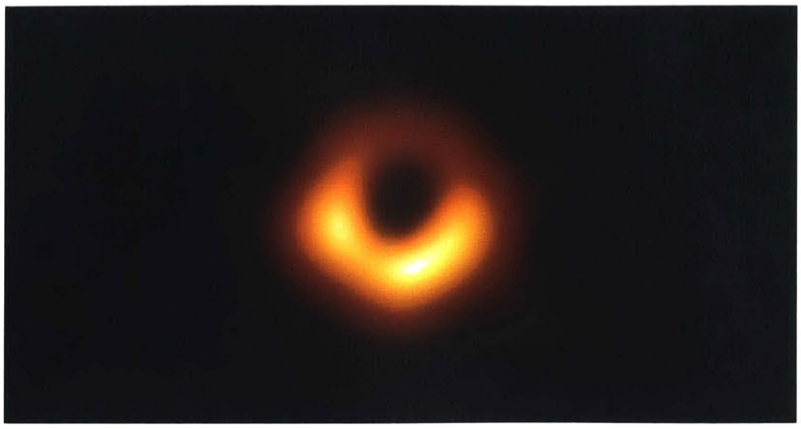


Figure 5-1: Black hole image from the center of the M87 galaxy.

argued that her contributions were overstated, since her collaborator Andrew Chael wrote the majority of the lines of code in the GitHub repository of the code. While it is incredibly simplistic to assign credit based on lines of code, this argument struck a chord, and led to a widespread harassment campaign againt Bouman.

At the heart of this controversy is the interaction between the complex sociotechnical networks that enabled the image, and people's psychology. Indeed the EHT team had over 200 people involved, many of whom played a crucial role in creating the final image. And some people couldn't stand the idea of a women getting a large portion of the credit. As our scientific, artistic and social world becomes more diffuse, networked and complicated, our ability to discern the correct allocation of credit may drop to zero (if one even exists at all). Add to this the extent to which our psychological biases and priors can motivate us to allocate responsibility in a certain way, and we are left in a troubling situation indeed. This thesis proposes using complex systems thinking to tackle these complex issues of responsibility and agency from on, and calls for further empirical research on how our psychological biases can impact our decision-making for these increasingly common situations.

5.2 On Agency

For the regime of AI systems, which have been incredibly disruptive and impactful in recent years, there is a particular psychological mechanism that is understudied:

the extent to which we endow agency to complex, computational systems. This phenomenon has been known since the 1960's with the coining of the term the ELIZA effect, but has remained mostly an abstract principle. Now, the proliferation of AI systems in applied settings has created a world where AI are entangled within social situations. Self-driving cars can murder pedestrians. Newsfeed algorithms can drive echo chambers. Disease detection algorithms can save people's lives. GANs can generate expensive paintings. Whether we like it or not, these AI systems have become relevant entities in our world, which requires us to apply our psychological machinery and heuristics of morality, responsibility and justice to them.

Anthropormorphicity is one such way that this agency manifests. When we endow our machines with more human-likeness, it changes the way we interact with them, and how we assign responsibility in these complicated sociotechnical situations. In order to have a more nuanced and adapted moral framework within these situations, we must further understand to what extent we anthropomorphize AI, and how that affects outcomes. When a programmer that makes risk recidivism software is critiqued for biased predictions, they currently might say, "it's not my fault, the algorithm is racist!" In this way, they deflect blame onto the AI, whose endowed agency allows it to "absorb the responsibility." Indeed our results suggest that increasing perceived anthropomorphicity does indeed causally reduce the allocation of responsibility to the practitioner, which is the first empirical demonstration of this idea. Thus, we must be careful about how we frame these algorithms, and how our psychology can be manipulated to prevent humans from being held accountable for problematic outcomes.

Conditional on this understanding, there is a second layer of inquiry, that involves the possibility of incentive-shaping. If people begin to understand that the way we talk about AI can indeed cause it to absorb responsibility, then that may induce them to behave in non-optimal ways. For example, in a complicated court case, a judge might overly rely on the output of an algorithm in an attempt to diminish their own responsibility if the criminal reoffends (again the hallow echo of "it's not my fault! It was the algorithm that messed up!") [118, 11]. People may even adapt to norms within self-driving cars by driving in such a way as to ensure that the algorithm is

actually blamed in the event of a crash [11]. As David Gelernter said, "the real danger is not machines becoming humans, rather but humans becoming machines [39]."

Bibliography

[1] Acm conference on fairness, accountability, and transparency (acm fat*). https://fatconference.org/, July 2018.

[2] Fat/ml. https://www.fatml.org, July 2018.

[3] Is artificial intelligence set to become art's next medium? https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx, Dec 2018.

[4] Neurips 2018 highlights (part 1). https://medium.com/dair-ai/neurips-2018-highlights-part-1-4814a39e0e80, Dec 2018.

[5] Students develop a system AI spirits that AI automatically synthesizes ghosts into any irrelevant image. https://gigazine.net/gsc_news/en/20181101-add-ghost-photo-ai-spirits/, Nov 2018.

[6] Better language models and their implications. https://openai.com/blog/better-language-models/, Feb 2019.

[7] Laura M Ahearn. Language and agency. *Annual review of anthropology*, 30(1):109–137, 2001.

[8] Ali Alkhatib and Michael Bernstein. Street–level algorithms: A theory at the gaps between policy and decisions. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2019.

[9] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23, 2016.

[10] Oxman Neri Antonelli, Paola and Kevin Slavin. Knotty objects. https://medium.com/mit-media-lab/the-summit-9a632339f56c, July 2015.

[11] Edmond Awad, Sydney Levine, Max Kleiman-Weiner, Sohan Dsouza, Josh Tenenbaum, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. Blaming humans in autonomous vehicle accidents: Shared responsibility across levels of automation. *arXiv preprint arXiv:1803.07170*, 2018.

[12] Amarabha Banerjee. Is youtube's AI algorithm evil? https://hub.packtpub.com/is-youtubes-ai-algorithm-evil/, Sep 2018.

[13] Robbie Barat. art-dcgan. https://github.com/robbiebarrat/art-DCGAN, Aug 2017.

[14] Solon Barocas, Sophie Hood, and Malte Ziewitz. Governing algorithms: A provocation piece. *Available at SSRN 2245322*, 2013.

[15] Gregory Bateson. *Steps to an ecology of mind: Collected essays in anthropology, psychiatry, evolution, and epistemology.* University of Chicago Press, 2000.

[16] Yoshua Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2017.

[17] James O Berger and Luis R Pericchi. The intrinsic bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122, 1996.

[18] Waltz Binaire. Used another AI called "content based fill" (photoshop) to enhance edmond belamy. https://twitter.com/WaltzBinaire/status/1058328182741450754, Nov 2018.

[19] Elizabeth Broadbent, Vinayak Kumar, Xingyan Li, John Sollers 3rd, Rebecca Q Stafford, Bruce A MacDonald, and Daniel M Wegner. Robots with display screens: a robot with a more humanlike face display is perceived to have more mind and a better personality. *PLOS ONE*, 8(8):e72589, 2013.

[20] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[21] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[22] Judee K Burgoon, Joseph A Bonito, Bjorn Bengtsson, Carl Cederberg, Magnus Lundeberg, and Lisa Allspach. Interactivity in human–computer interaction: A study of credibility, understanding, and influence. *Computers in human behavior*, 16(6):553–574, 2000.

[23] Alex Campolo, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford. AI now 2017 report. *AI Now Institute at New York University*, 2017.

[24] Shan Carter and Michael Nielsen. Using artificial intelligence to augment human intelligence. *Distill*, 2(12):e9, 2017.

[25] Álvaro Castro-González, Henny Admoni, and Brian Scassellati. Effects of form and motion on judgments of social robots animacy, likability, trustworthiness and unpleasantness. *International Journal of Human-Computer Studies*, 90:27–38, 2016.

[26] François Chollet. Ganism (the specific look and feel of seemingly gan-generated images) may yet become a significant modern art trend. https://twitter.com/fchollet/status/885378870848901120, Jul 2017.

[27] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148, 2018.

[28] Christopher W Clifton, Deirdre K Mulligan, and Raghu Ramakrishnan. Data mining and privacy: An overview. In *Privacy and Technologies of identity*, pages 191–208. Springer, 2006.

[29] Gabe Cohn. AI art at christie's sells for 432,500. https://www.nytimes.com/2018/10/25/arts/design/ai-art-sold-christies.html, Oct 2018.

[30] Ahmed Elgammal. What the art world is failing to grasp about christie's AI portrait coup. https://www.artsy.net/article/artsy-editorial-art-failing-grasp-christies-ai-portrait-coup, Oct 2018.

[31] Madeleine Elish. When your self-driving car crashes, you could still be the one who gets sued. https://qz.com/461905/when-your-self-driving-car-crashes-you-could-still-be-the-one-who-gets-sued/, July 2015.

[32] MC Elish. Moral crumple zones: Cautionary tales in human-robot interaction (we robot 2016). 2016.

[33] Nicholas Epley, Adam Waytz, and John T Cacioppo. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4):864, 2007.

[34] Ziv Epstein, Blakeley H Payne, Judy Hanwen Shen, Abhimanyu Dubey, Bjarke Felbo, Matthew Groh, Nick Obradovich, Manuel Cebrián, and Iyad Rahwan. Closing the AI knowledge gap. *arXiv preprint arXiv:1803.07233*, 2018.

[35] Jessica Feld. Template license and collaboration agreements for AI art. http://blogs.harvard.edu/cyberlawclinic/2019/02/04/template-license-and-collaboration-agreements-for-ai-art/, Feb 2019.

[36] Julia Fink. Anthropomorphism and human likeness in the design of robots and human-robot interaction. In *International Conference on Social Robotics*, pages 199–208. Springer, 2012.

[37] Agence France-Presse. Portrait made entirely using AI algorithm sells for more than 400,000. https://www.ndtv.com/world-news/

edmond-de-belamy-made-entirely-using-ai-algorithm-sells-at-christies-for-more-Oct 2018.

[38] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347, 1996.

[39] David Gelernter. The danger is not machines becoming humans, but humans becoming machines. https://bigthink.com/in-their-own-words/the-danger-is-not-machines-becoming-humans-but-humans-becoming-machines, Dec 2013.

[40] Will M Gervais. Perceiving minds and gods: How mind perception enables, constrains, and is triggered by belief in gods. *Perspectives on Psychological Science*, 8(4):380–394, 2013.

[41] Barbara Goldberg. First-ever auction of AI-created artwork set for christie's gave. https://www.reuters.com/article/us-france-art-artificial-intelligence/first-ever-auction-of-ai-created-artwork-set-for-christies-gavel-idUSKCN1MX2WO Oct 2018.

[42] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[43] Heather M Gray, Kurt Gray, and Daniel M Wegner. Dimensions of mind perception. *science*, 315(5812):619–619, 2007.

[44] Kurt Gray, Liane Young, and Adam Waytz. Mind perception is the essence of morality. *Psychological inquiry*, 23(2):101–124, 2012.

[45] Matt Groh. Deep angel. https://deepangel.media.mit.edu, Aug 2018.

[46] Kerstin S Haring, David Silvera-Tawil, Tomotaka Takahashi, Katsumi Watanabe, and Mari Velonaki. How people perceive different robot types: A direct comparison of an android, humanoid, and non-biomimetic robot. In *2016 8th International Conference on Knowledge and Smart Technology (KST)*, pages 265–270. IEEE, 2016.

[47] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[48] Leonhard Held. Introducing bayes factors. 2011.

[49] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701, 2015.

[50] Ralph Hertwig and Ulrich Hoffrage. *Simple heuristics in a social world.* Oxford University Press, 2013.

[51] Aaron Hertzmann. Can computers create art? In *Arts*, volume 7, page 18. Multidisciplinary Digital Publishing Institute, 2018.

[52] Aaron Hertzmann. Who should get the credit for AI articlet? `http://www.cnn.com/style/article/ai-art-who-should-get-credit-conversation/?iid=ob_article_footer_expansion`, Apr 2019.

[53] Anthony JG Hey, Stewart Tansley, Kristin M Tolle, et al. *The fourth paradigm: data-intensive scientific discovery*, volume 1. Microsoft research Redmond, WA, 2009.

[54] Natashah Hitti. Christie's sells AI-created artwork painted using algorithm for 432,000. `https://www.dezeen.com/2018/10/29/christies-ai-artwork-obvious-portrait-edmond-de-belamy-design/`, Oct 2018.

[55] Douglas R Hofstadter. *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought.* Basic books, 1995.

[56] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudík, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? *arXiv preprint arXiv:1812.05239*, 2018.

[57] John J Horton, David G Rand, and Richard J Zeckhauser. The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, 14(3):399–425, 2011.

[58] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[59] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[60] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[61] William Joseph King. Anthropomorphic agents: Friend, foe, or folly. 1995.

[62] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[63] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

[64] Jonathan F Kominsky, Jonathan Phillips, Tobias Gerstenberg, David Lagnado, and Joshua Knobe. Causal superseding. *Cognition*, 137:196–209, 2015.

[65] Sören Krach, Frank Hegel, Britta Wrede, Gerhard Sagerer, Ferdinand Binkofski, and Tilo Kircher. Can machines think? interaction and perspective taking with robots investigated via fmri. *PLOS ONE*, 3(7):e2597, 2008.

[66] Nicole C Krämer, Astrid von der Pütten, and Sabrina Eimler. Human-agent and human-robot interaction theory: similarities to and differences from human-human interaction. In *Human-computer interaction: The agency perspective*, pages 215–240. Springer, 2012.

[67] David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. Computational social science. *Science*, 323(5915):721–723, 2009.

[68] Kalev Leetaru. Does AI truly learn and why we need to stop overhyping deep learning. `https://www.forbes.com/sites/kalevleetaru/2018/12/15/does-ai-truly-learn-and-why-we-need-to-stop-overhyping-deep-learning/#58f3435d68c0`, Dec 2018.

[69] Zachary C Lipton, Kamyar Azizzadenesheli, Abhishek Kumar, Lihong Li, Jianfeng Gao, and Li Deng. Combating reinforcement learning's Sisyphean curse with intrinsic fear. *arXiv preprint arXiv:1611.01211*, 2016.

[70] Zachary C Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship. *arXiv preprint arXiv:1807.03341*, 2018.

[71] R Duncan Luce. *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2012.

[72] Bertram F Malle. Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics and Information Technology*, 18(4):243–256, 2016.

[73] Aarian Marshall. After a deadly crash, uber returns robocars to the road. `https://www.wired.com/story/uber-returns-self-driving-after-deadly-crash/`, Dec 2018.

[74] Molly C Martini, Christian A Gonzalez, and Eva Wiese. Seeing minds in others–can agents with robotic appearance have human-like preferences? *PLOS ONE*, 11(1):e0146310, 2016.

[75] Maya B Mathur and David B Reichling. Navigating a social world with robot partners: A quantitative cartography of the uncanny valley. *Cognition*, 146:22–32, 2016.

[76] Lois McNay. *Gender and agency: Reconfiguring the subject in feminist and social theory.* John Wiley & Sons, 2013.

[77] Brett Molina. Christie's sells painting created by artificial intelligence for 432,500. `https://www.usatoday.com/story/news/nation-now/2018/10/25/painting-created-ai-going-auction-block-christies/1759967002/`, Oct 2018.

[78] Michael C Mozer. Neural network music composition by prediction: Exploring the benefits of psychoacoustic constraints and multi-scale processing. *Connection Science*, 6(2-3):247–280, 1994.

[79] Ritesh Noothigattu, Snehalkumar S Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D Procaccia. A voting-based system for ethical decision making. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[80] Ara Norenzayan, Will M Gervais, and Kali H Trzesniewski. Mentalizing deficits constrain belief in a personal god. *PLOS ONE*, 7(5):e36880, 2012.

[81] Obvious. Obvious, explained., Feb 2018.

[82] Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Broadway Books, 2017.

[83] Robert Trappl Paolo Petta Sabine Payr. *Emotions in humans and artifacts.* MIT Press, 2002.

[84] David Pescovitz. An artificial intelligence populated these photos with glitchy humanoid ghosts. `https://boingboing.net/2018/10/31/an-artificial-intelligence-pop.html`, Oct 2018.

[85] Elizabeth Phillips, Xuan Zhao, Daniel Ullman, and Bertram F Malle. What is human-like?: Decomposing robots' human-like appearance using the anthropomorphic robot (abot) database. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 105–113. ACM, 2018.

[86] Jonathan Phillips and Jonathan F Kominsky. Causation and norms of proper functioning: Counterfactuals are (still) relevant. In *CogSci*, 2017.

[87] Aaron Powers and Sara Kiesler. The advisor robot: tracing people's mental model from a robot's physical attributes. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 218–225. ACM, 2006.

[88] Diane Proudfoot. Anthropomorphism and AI: Turing's much misunderstood imitation game. *Artificial Intelligence*, 175(5-6):950–957, 2011.

[89] Iyad Rahwan. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*, 20(1):5–14, 2018.

[90] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. Machine behaviour. *Nature*, 568(7753):477, 2019.

[91] David G Rand, Alexander Peysakhovich, Gordon T Kraft-Todd, George E Newman, Owen Wurzbacher, Martin A Nowak, and Joshua D Greene. Social heuristics shape intuitive cooperation. *Nature communications*, 5:3677, 2014.

[92] Laurel D Riek, Tal-Chen Rabinowitch, Bhismadev Chakrabarti, and Peter Robinson. How anthropomorphism affects empathy toward robots. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 245–246. ACM, 2009.

[93] Steve Rousseau. This AI inserts unsettling digital ghosts into normal pictures. `http://digg.com/2018/ai-spirits-mit`, Oct 2018.

[94] Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991.

[95] Sarag Schwettman. Generist project resources + examples. `https://github.com/schwettmann/generism`, Feb 2019.

[96] Nick Seaver. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society*, 4(2):2053951717738104, 2017.

[97] Andrew D Selbst, Sorelle Friedler, Suresh Venkatasubramanian, Janet Vertesi, et al. Fairness and abstraction in sociotechnical systems. In *ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, 2018.

[98] Azim F Shariff, Joshua D Greene, Johan C Karremans, Jamie B Luguri, Cory J Clark, Jonathan W Schooler, Roy F Baumeister, and Kathleen D Vohs. Free will and punishment: A mechanistic view of human nature reduces retribution. *Psychological science*, 25(8):1563–1570, 2014.

[99] Elke Sharz. This is why the moral machine project is so problematic: this is not ethical thinking. these are the statistical results for a survey sourced via an online gaming interface based on binary choices. we must not confuse this with appropriate ethical deliberation for tech. please. `https://twitter.com/elkeschwarz/status/1078536625955557376?s=21`, Dec 2018.

[100] Tom Simonite. A 'neurographer' puts the art in artificial intelligence. `https://www.wired.com/story/neurographer-puts-the-art-in-artificial-intelligence/`, July 2017.

[101] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[102] Valerie K Sims, Matthew G Chin, David J Sushil, Daniel J Barber, Tatiana Ballion, Bryan R Clark, Keith A Garfield, Michael J Dolezal, Randall Shumaker, and Neal Finkelstein. Anthropomorphism of robotic forms: A response to affordances? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 49, pages 602–605. SAGE Publications Sage CA: Los Angeles, CA, 2005.

[103] Adam Smith. Christie's to auction art created by artificial intelligence. `https://www.pcmag.com/news/364580/christies-to-auction-art-created-by-artificial-intelligence`, Oct 2018.

[104] Ted Striphas. Algorithmic culture. *European Journal of Cultural Studies*, 18(4-5):395–412, 2015.

[105] TiredOldCrow. Irresponsible anthropomorphism is killing AI journalism. `https://www.reddit.com/r/MachineLearning/comments/b0rdsi/d_irresponsible_anthropomorphism_is_killing_ai/`, March 2019.

[106] Amos Tversky. Choice by elimination. *Journal of mathematical psychology*, 9(4):341–367, 1972.

[107] Amos Tversky. Elimination by aspects: A theory of choice. *Psychological review*, 79(4):281, 1972.

[108] D Ulyanov, A Vedaldi, and V Lempitsky. Instance normalization: the missing ingredient for fast stylization. cscv. *arXiv preprint arXiv:1607.08022*, 2017.

[109] James Vincent. How three french students used borrowed code to put the first AI portrait in christie's. `https://www.theverge.com/2018/10/23/18013190/ai-art-portrait-auction-christies-belamy-obvious-robbie-barrat-gans`, Oct 2018.

[110] Lev S Vygotsky. Thinking and speech. *The collected works of LS Vygotsky*, 1:39–285, 1987.

[111] Lev Semenovich Vygotsky. *Mind in society: The development of higher psychological processes*. Harvard university press, 1980.

[112] Jane Wakefield. Are you scared yet? meet norman, the psychopathic AI. `https://www.bbc.com/news/technology-44040008`, June 2018.

[113] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 5, 2018.

[114] Adam Waytz, John Cacioppo, and Nicholas Epley. Who sees human? the stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, 5(3):219–232, 2010.

[115] Adam Waytz, Joy Heafner, and Nicholas Epley. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52:113–117, 2014.

[116] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

[117] James V Wertsch, Peter Tulviste, and Fran Hagstrom. A sociocultural approach to agency. *Contexts for learning: Sociocultural dynamics in children's development*, 23:336–356, 1993.

[118] Eric Westervelt. Did a bail reform algorithm contribute to this san francisco man's murder? `https://www.npr.org/2017/08/18/543976003/did-a-bail-reform-algorithm-contribute-to-this-san-francisco-man-s-murder`, Aug 2018.

[119] Aiyana K Willard and Ara Norenzayan. Cognitive biases explain religious belief, paranormal belief, and belief in life's purpose. *Cognition*, 129(2):379–391, 2013.

[120] Raymond Williams. *Culture and society, 1780-1950*. Columbia University Press, 1983.

[121] Mark Wilson. AI is making halloween so much spookier. `https://www.fastcompany.com/90258225/ai-is-making-halloween-so-much-spookier`, Oct 2018.

[122] Pinar Yinardag, Manuel Cebrian, and Iyad Rahwan. Norman. world's first psychopath AI. `http://norman-ai.mit.edu/`, April 2018.

[123] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. *arXiv preprint arXiv:1801.07892*, 2018.

[124] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[125] Jichen Zhu and D Fox Harrell. The artificial intelligence (AI) hermeneutic network: A new approach to analysis and design of intentional systems. In *Proceedings of the 2009 Digital Humanities Conference*, pages 301–304, 2009.