Mimetic Models: Ethical Implications of AI that Acts Like You

Reid McIlroy-Young
reidmcy@cs.toronto.edu
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada

Jon Kleinberg
Department of Computer Science
Cornell University
Ithica, New York, USA

Siddhartha Sen
Microsoft Research
New York City, New York, USA

Solon Barocas
Microsoft Research
& Cornell University
New York City, New York, USA

Ashton Anderson
Department of Computer Science
University of Toronto
Toronto, Ontario, Canada

ABSTRACT

An emerging theme in artificial intelligence research is the creation of models to simulate the decisions and behavior of specific people, in domains including game-playing, text generation, and artistic expression. These models go beyond earlier approaches in the way they are tailored to individuals, and the way they are designed for interaction rather than simply the reproduction of fixed, pre-computed behaviors. We refer to these as *mimetic models*, and in this paper we develop a framework for characterizing the ethical and social issues raised by their growing availability. Our framework includes a number of distinct scenarios for the use of such models, and considers the impacts on a range of different participants, including the target being modeled, the operator who deploys the model, and the entities that interact with it.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing devices**; • **Computing methodologies** → **Artificial intelligence**; • **Social and professional topics** → **Computing / technology policy**.

KEYWORDS

Artificial Intelligence; Machine Learning; Generative Models; Mimetic Models; Ethics

ACM Reference Format:
Reid McIlroy-Young, Jon Kleinberg, Siddhartha Sen, Solon Barocas, and Ashton Anderson. 2022. Mimetic Models: Ethical Implications of AI that Acts Like You. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3514094.3534177

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08…$15.00
https://doi.org/10.1145/3514094.3534177

1 INTRODUCTION

When machine learning (ML) is deployed to replace human effort on tasks in specific application domains, the primary focus has traditionally been on the performance of the ML system relative to human capability on the relevant tasks [44, 68, 69, 91, 92]; but there has been increasing interest in trying to design ML solutions that exhibit human-like behavior on the task, generating solutions that look like what a skilled human being would produce [6, 43, 58, 63]. In domains where there is extensive data on individual behavior, it becomes possible to build such models not simply on aggregate human behavior, but tailored to the behavior of *specific (individual) people*—a model that tries to simulate the actions of a particular person in arbitrary situations within the domain.

The idea of designing ML models to simulate specific people is becoming a reality in a growing number of domains—particularly for game-playing, where chess engines have been trained to play like specific human chess players [64], and e-sports avatars have been trained to play like specific human athletes [71, 97, 98]; and for writing and text generation, where models have been trained to produce text in the writing style of specific authors [8] or social media users [86]. From these realized examples, it becomes possible to see how the same techniques could be used in other forms of artistic expression (for example, to compose music in the style of specific people), or professional expertise (where work in medical AI is beginning to explore the design of models that try to match the diagnoses of specific doctors [36]). The concreteness of these developments makes clear that it is an appropriate time to identify the common themes across the efforts in these different domains, and to consider their ethical and social implications.

With this in mind, we define a *mimetic model* to be an algorithm that is trained on data from a *specific individual* in a given domain, and which is designed to accurately predict and *simulate* the behavior of this individual in *new situations* from the domain. This definition is intended to capture the examples discussed above, and to highlight the key themes that we believe are central to them. Crucially, a mimetic model is *generative* in the sense that it does not exist simply to predict a specific person's behavior, but to produce this behavior in context and thus interact with new environments and new individuals. In this way, the mimetic model is broader than any one of its outputs; it is not simply an e-mail message, tweet, or chess move that looks like it was created by you, but a mechanism that can be placed in arbitrary situations and produce messages,

tweets, or chess moves that are designed to resemble what you would do in these situations.

1.1 Mimetic Models: Analyzing their Ethical and Social Implications

Within this framework, we ask: What normative issues come into play when mimetic models enter more widespread use across a diverse set of domains? In posing this question, we note that the normative impacts can be both produced and experienced by several different parties: the *target* individual that the mimetic model is designed to simulate; the *creator* who builds the model and the *operator* who uses it; and finally the party who *interacts* with the model.

One of the most immediate concerns about a mimetic model is its potential to be used for deception: someone could believe that they are interacting with you when they are interacting with a mimetic model of you. These concerns are related to the role that deepfakes play in spreading misinformation [102], and we discuss the relationships and distinctions with mimetic models further in Section 4.2.1.

But an animating motivation of the present paper is to realize how many additional normative concerns remain even when a mimetic model is not being used deceptively—that is, even when there is transparency about which agents in a given domain are mimetic models, and which target individuals they are based on. We organize these concerns into three broad categories, which we analyze in the subsequent sections of the paper. To give a sense for some of the questions that motivate this analysis, we begin with an overview of the three categories and some of the issues that arise in each.

Mimetic Models as a Means. First, interacting with a mimetic model can be used as preparation for interactions in real life—essentially, as a means to an end, where the end is the real-life interaction. For example, in a setting where two people A and B are going to meet so that B can interview A for a job, we could imagine that A might "practice" the interview dozens of times with a mimetic model of B so as to find the types of answers that seem most appealing to B. And in place of a job interview, we could imagine that A is repeatedly practicing their interactions with B in preparation for a journalistic interview, for a fund-raising pitch, or for going on a date. You might feel differently, for example, being interviewed by a journalist if you knew they had spent several days practicing their interview on a mimetic model of you. And in a competition such as a chess game, where we've noted that mimetic models are already feasible, A could practice against a mimetic model of a future opponent B to identify their weaknesses.

We might reasonably feel that there are qualitative contrasts between the way this type of preparation operates across different domains—for example, rehearsing a chess game against a mimetic model may raise different concerns than rehearsing a conversation with a future date. But in all cases, the availability of mimetic models would change the underlying norms and expectations that we would have for one-on-one interactions between people, the expectations for what people should disclose about the context of these interactions, and the conclusions we draw from them as a result.

Mimetic Models as an End. Another broad category of uses we envision are those in which a mimetic model is used not to prepare for future interactions, but as an end in itself. In some cases, this end might be the potential for interaction with the mimetic model. For example, chatbots have been used as automated teaching assistants in online class forums [15], but these have tended to be automated agents trained on aggregate data. We could ask how the normative considerations might change if an online class created a mimetic model of each human TA, so that students who preferred the style of office-hour help offered by human TA A could choose their mimetic model over the model of human TA B. This is a form of work replacement through automation that is highly personalized, and raises questions about how human labor can become devalued, about the forms of consent and compensation the real A and B are entitled to, and about the responsibilities the model creators have to faithfully represent the behaviors of A and B, given that will be presented to the world through their mimetic models. If the mimetic model of A is a highly accurate representation but also rude to students, should that reflect poorly on the real A? If students develop a social dynamic in which they are verbally abusive on the forum to the mimetic model of A, what harms has the real A suffered as a result?

Mimetic models can also be used as an end in themselves when they are deployed to satisfy the interest of a group that functions as an audience or a set of spectators. For example, "what if" scenarios are a source of fascination among fans of all sports—in a given sport, what if A, the greatest player of the 1980s, had had the opportunity to play at their peak against B, the greatest player of the present day? Mimetic models provide a new mechanism for such thought experiments, and they raise analogous questions of reputation, compensation, and consent. If a mimetic model of former chess champion A defeats a mimetic model of former chess champion B, what do we expect to be the reputational consequences in practice for the real A and B? And what stake should A and B have in the creation and use of their mimetic models for these purposes? More narrowly scoped precursor questions have been the subject of litigation in the video-game domain, when statistics and likenesses of athletes have been used without consent or compensation [19]; the increasing fidelity of mimetic models has the potential to intensify all of these considerations.

Mimetic Models of Oneself. Many of the normative considerations involving the use of mimetic models depend on the relationship of the model's target to its creator and/or operator. The cases above make clear how diverse this set of relationships can be. But a relationship that brings up a specific set of considerations, and therefore benefits from separate analysis, is the case in which the operator of the mimetic model is its target—that is, the case in which someone builds a mimetic model of themself.

There are several natural uses for a mimetic model of oneself. One of the most basic is as a *force multiplier*; for example, a model trained to generate e-mail replies in your style could be used to answer significantly more messages than you are practically able to handle on your own. This raises questions about the level of

disclosure that is appropriate for the authorship of such messages, and how norms about appropriateness will evolve. A world in which e-mail messages written by a mimetic model are explicitly flagged as such may produce different social cues than a world in which it is left ambiguous which messages were authored by the real you and which by your mimetic model. The use of effort as a means of signaling commitment to a relationship would look different in these two worlds. And even under a norm where people explicitly flag messages that were written by their model, there are more subtle choices about what to disclose and what to reveal. Is it deceptive, for example, for you to use a mimetic model whose level of politeness has been covertly increased to a level beyond your own natural politeness? Is this fundamentally more deceptive than manually following advice, without the use of a model, for how to write e-mail that sounds more polite?

1.2 Framework and Related Concepts

With this range of potential scenarios in mind, it is useful to return to the general properties that characterize a mimetic model. As discussed above, there are three crucial aspects to this type of model: (i) it is targeted to a specific individual, rather than attempting to simulate human behavior in an aggregate sense; (ii) it is generative, in that it produces new behaviors; and (iii) it is interactive, in that it generates these behaviors in response to interactions with other individuals or with new environments.

This structure, as well as the scenarios above, make clear that there are four roles that are important in any mimetic model:

- The *target*, whose behavior the mimetic model is designed to simulate. We will say that the *fidelity* of the mimetic model is its accuracy in matching the behavior of the target.
- The *creator*, who builds the mimetic model. This implies that, at some level, the creator has at least indirect access to data about the target.
- The *operator*, who uses the mimetic model.
- The *interactor*, who engages in some form of interaction with the mimetic model. In different scenarios, the interactor might be communicating with the mimetic model, competing against it, or potentially watching it as a spectator.

Because AI systems interact with human behavior in so many different ways, it is also useful to situate the notion of mimetic models in comparison to related concepts. Of course, different concepts will naturally blend into each other, and so some of the distinctions that we draw here are questions of degree rather than absolute contrasts.

First, *recommendation systems* naturally depend on personalized models of their users [9, 85]. We think of these as distinct from the general formulation of mimetic models in that user models for recommendations tend to be focused on the narrow task of predicting a user's preferences for particular pieces of content, and providing content that is likely to satisfy the user.

In a different direction, *deepfakes* are a type of manipulated media—often video—designed to portray specific people engaging in behaviors that didn't occur in real life [101]. These can be used in deceptive or defamatory ways, or in instances where for example a deepfake of an actor's image or voice is used in a movie where they could not appear [4]. Deepfakes clearly raise a number of ethical considerations that parallel what we consider for mimetic models, but it is important to note the key distinction that deepfakes tend not to be designed for unrestricted interaction with their environment or with others, but rather to present a static, precomputed set of behaviors.

We discuss these comparisons, as well as additional related concepts, in Section 4.2 later in the paper. We turn next to a more in-depth discussion of our main categories of scenarios. Throughout our analysis, we focus on characterizing the novel ethical and social questions that mimetic models raise. Fully addressing these questions will likely involve a large collective effort over a number of years.

2 APPLICATIONS OF MIMETIC MODELS

One way to classify the potential applications of mimetic models is to start by considering the possible ways through which an event in the world might be affected by the existence of such a model. At a high level, the event might be affected because an individual arrives at the event better prepared through their prior interaction with a mimetic model; or the event might be affected because the mimetic model directly participates in the event.

In the former case, we think of the mimetic model as a means to an end, in that it prepares someone for a future interaction but is not necessarily present when the event takes place. For example, a mimetic model might be used to help people prepare to interact with the actual person who is the target of the model. In this case, mimetic models would serve as a way for people to learn how best to achieve their goals in interacting with a person by first interacting with the mimetic model of the person.

In the latter case, when the mimetic model directly participates in the event, such a model could potentially be used as a complete substitute for the person who is the target of the model. For such cases, we make a further distinction between (i) scenarios in which an event that could have occurred with the genuine target instead takes place with the model, and (ii) counterfactual scenarios that could not feasibly have occurred without the presence of a mimetic model: for example, scenarios in which mimetic models of athletes or artists from different eras interact with one another—a type of interaction that could not have happened in real life.

Across all of these scenarios, we also consider the special case where the target and operator of a mimetic model is the same person. In such cases, a person might use a model of themself as a means to an end, having the model explore the world on the person's behalf to help the person better prepare to act in it themself; and as an end in itself, offloading certain tasks that the person would have otherwise needed to perform themself.

To give each of these possibilities greater substance, we consider a range of more concrete scenarios that illustrate how this might work, with some scenarios already visible in practice, others practically feasible, and still others being possibly feasible in the (perhaps distant) future. In progressing through these different scenarios, we hope to highlight the different ethical issues that different uses of mimetic models might raise. We organize the section based on the distinctions discussed above, beginning with mimetic models as a means to an end (Section 2.1), then as an end in themselves

(Section 2.2), and finally for the case in which an individual creates a model of themself (Section 2.3).

2.1 Mimetic Modelling as a Means to an End

We first consider the ways in which mimetic models might be used as a means to an end—that is, as a way to learn about the target of the mimetic model so as to be better able to achieve certain goals when interacting with the actual person in the future.

Preparing for a competition. Imagine a person who has access to a mimetic model of a future opponent that they hope to defeat in an upcoming chess tournament. Further imagine that the person can rely on the mimetic model of their opponent to see how the opponent would respond to different moves and strategies. For example, to prepare to play the opponent at the tournament, the person could play as many games against the mimetic model as time allows. The person could also see how the mimetic model would respond to specific positions, rather than playing a full game linearly to its conclusion. Or the person could make a move, see how the mimetic model responds, and, if the move did not have the anticipated benefit, take back the move to try an alternative to see if that would be any more successful. The person could even have super-human (i.e., non-mimetic) chess-playing agents play against the mimetic model of their opponent to discover weaknesses that the person would not have even thought to test for.

Such a scenario is not fantastical. Recent research has demonstrated that it is possible to build mimetic models of particular players when there are available records of people's past game play [6, 23, 34, 43, 63, 64, 68, 71, 92, 97, 98, 100], whether we're considering Chess [100], Go [68], Shogi [92], Hanabi [6], Diplomacy [43], or other games with a finite set of legal moves. In these games, player actions can be recorded with perfect accuracy. Relying on players' past games as training data, it is thus possible to create a deep learning-based model that would likely make the moves of specific players. Recent work shows that building such player-specific models is even possible with a rather small sample of a player's past games [64].

What ethical issues does such a scenario raise? In particular, what, if anything, is different about a person preparing to play an opponent by looking over the opponent's publicly available past game play, which is common practice in competitive chess, and playing a mimetic model of the opponent? What advantage, if any, does the mimetic model give the person preparing for this match in comparison to the more traditional ways that a person might prepare?* One way to try to answer this question is to compare how the person learns under these two different scenarios. When a person is trying to learn from an opponent's past game play, they must expend considerable effort reviewing all of their opponent's past game play and attempt to generalize from these examples—that is, to not only memorize how the opponent has acted in the face of specific positions, but to induce a rule from past game play that would indicate how the opponent would act in the face of previously unencountered positions. Reliably extrapolating from an opponent's past game play is a non-trivial task both in terms of the time that must be invested by the person and the cognitive demands placed on them. A mimetic model would essentially do this work for the person: it would generalize from the opponent's past game play, relieving the person of the burden of manually studying individual games and positions, while also likely exceeding the person's ability to generalize accurately from these past examples. Note, however, that while the mimetic model might have assumed these burdens, the person still needs a way to learn from the lessons that the mimetic model has drawn from the opponent's past game play. The obvious way that the person might try to do this is to play games against the mimetic model or see how the mimetic model responds to specific positions, as described above. This then raises the question of whether learning about an opponent by playing a mimetic model of them is a more effective or efficient way to prepare for playing them than simply reviewing the opponent's past game play. As mentioned, there is good reason to believe the mimetic models will be able to generalize more accurately from opponent's past game play than humans. Indeed, the value of machine learning in many settings is that it can detect patterns and signals that go overlooked by humans. Yet it is still an open empirical question if playing a mimetic model offers meaningful advantages over traditional training methods.

If it turns out that mimetic models enhance a person's ability to prepare to play an opponent, then mimetic models have obvious implications for fair competition, especially if mimetic models are not universally available. We might be less concerned with such a development if the opponent that the person is preparing to play also had a mimetic model of the person to train against. But if only one of the two opponents has access to a mimetic model, then it poses an obvious threat to competition. While certain chess players might already benefit from access to resources and training that are not available to others, mimetic models could further exacerbate these disparities, eroding the equal playing field on which we hope players will compete.

Preparing for an interview. Consider a person about to undertake a job interview who happens to have access to a mimetic model of the person who will interview them. The interviewee might attempt to gain an edge on the interviewer by completing a round of test interviews with the mimetic model. In so doing, the interviewee might learn the specific things about themself that they would be wise to withhold and the specific things about themself that they would do well to highlight—that is, the interviewee might be able to figure out how to make the best possible impression, given what they have to offer and given what the interviewer is looking for. Access to a mimetic model of the interviewer could also allow the interviewee to test out different persuasive styles. Even when presenting the exact same facts about themself and their career, the interviewee might communicate these quite differently, with some presentations of these facts being much more compelling than others from the point of view of the interviewer. The interviewee might therefore test out a range of different approaches on the mimetic model, adopting a more aggressive and boastful style in one interaction before trying out a more agreeable and modest style in the next. The mimetic model could help the interviewee hone their tone to increase the likelihood that the interviewer will be left

*Participants having too much information about competitors' strategies (solving the 'metagame' [53, 70]) in a tournament is something that tournament operators already know to guard against [94], since it degrades the experience for participants and observers by reducing the diversity of strategies.

with a favorable impression. The interviewee could even rely on the mimetic model to learn personal details about the interviewer that would seem to have nothing to do with the job, but which might help the interviewee cultivate greater rapport with the interviewer. For example, the interviewee might learn that the interviewer is a baseball fan, that they own two dogs, and that they had a difficult divorce. The interviewee might try to establish some degree of affinity with the interviewer by strategically weaving these topics into the conversation, bonding over shared interests and gaining confidence by demonstrating sympathies for personal challenges.

Note that this scenario differs from the previous one insofar as the interaction is not zero-sum. In chess and other competitions, one person's gain is another person's loss: when a person learns the weakness of their opponent, the opponent necessarily suffers. The situation is different in the case of a job interview because there can be some alignment of interests. An interviewer might be pleased that the interviewee has communicated information about the characteristics of interest. Setting aside the possibility that an interviewee might simply lie about their qualifications or manufacture details that their interactions with the mimetic model suggest would impress the interviewer, there can be mutual benefits to an interviewee learning how best to interact with an interviewer. Of course, many of the things that the interviewee might learn about the interviewer via the mimetic model might be valuable not because they allow the interviewee to be assessed more accurately on their merits. Instead, the mimetic model might reveal personal qualities about the interviewer that the interviewee can exploit to compensate for their lack of merit. It's not obvious that an interviewer would be well served by someone who has simply figured out how to push their buttons.

Indeed, mimetic models could easily make people far more vulnerable to manipulation and exploitation. In everyday life, people rarely have the chance to try their luck multiple times to figure out the optimal steps to get what they want from an interaction. Learning intimate details about a person—their preferences and propensities, but also deeply private facts—often requires making yourself vulnerable to the person in the process. Their is some risk involved in feeling out an interviewer: they get to know something about you as you try to get to know something about them. Mimetic models undermine this symmetry.

Beyond interviews. While we've focused on interviews, such dynamics apply to a range of activities in which two parties are attempting to learn about and assess each other. As mentioned earlier, a mimetic model might help prepare for pitch meetings, but also interactions that seem much more distant from interviews.

Dating is a particularly useful scenario to contemplate because our instinctive reactions to using mimetic models in that context are normatively instructive. Imagine that A is going on a first date with B and hopes that it will lead to a longer-term relationship; and imagine that, as in our job-interview scenario, A prepares for the date by interacting with a mimetic model of B. There are some basic contrasts with the job-interview setting that may shift our normative assessment. In particular, a job interview is fundamentally transactional, and we evaluate the use of a mimetic model against the integrity of the transaction. In contrast, a first date is part of a potentially longer-term relationship that involves a range of other qualities, including establishing trust as a basis for intimacy, and the way in which this trust is established through expectations about the nature of the interaction.

We can also ask how the use of a mimetic model differs from other forms of preparation that A might do for their date with B, such as asking B's friend C for advice on what to emphasize in conversation. We have an intuitive sense that interaction with an actual model of B may be a qualitatively different type of preparation; this difference is reflected in pop culture's fascination with versions of this precise scenario, in the perfecting of repeated interactions in movies like *Groundhog Day*. Indeed, to have access to a mimetic model of someone begins to approximate the experience of being able to repeat a "time loop" with them. And this is a reflection of a point from earlier in this section, that the power of machine learning in general is to identify patterns that escape the unaided perception of human beings. In this way, the mimetic model of B may encode things about B that would be practically infeasible for A to discern on their own.

The dating scenario makes salient the lack of informed consent [22]. In this example and others, the mimetic-model-informed interaction is made more powerful by the target's lack of knowledge of how the model was used. Even if the model were trained on purely public data that the target was aware of, the model's (potentially superhuman) ability to provide specialized feedback in concrete situations raises a natural concern about whether this use requires consent. As the creator of a mimetic model often differs from the target, the question of consent persists through all of the scenarios we consider.

2.2 Mimetic Modelling as an End in Itself

In addition to being used indirectly to inform some future interaction, mimetic models could also be used directly as ends in their own right. In this Section, we explore a number of scenarios in which one's interaction with mimetic models is the end goal.

Target replacement. In many cases, people will be able to deploy mimetic models directly into important interactions. For example, imagine that an entrepreneur runs an online tutoring service, and employs a particularly popular and idiosyncratic tutor A. When parents inquire about the tutoring service, they most often wonder if A is available to teach their children. If the entrepreneur has access to a mimetic model of A, they could temporarily substitute the model for A when A is unable to work, for example if A is out sick. If the mimetic model satisfies customers just as well as A does, the entrepreneur may wonder if they still need A's services at all, and could opt to permanently replace A with the mimetic model of A. The entrepreneur may even go further, and wonder if the customer base as a whole would be more satisfied if everyone could be served by the model of A, rather than the various other human tutors under their employment. As another example, imagine that the reigning chess world champion Magnus Carlsen is not available to play in the online tournament you are organizing. You could opt to substitute the mimetic model of Magnus so that the other participants and viewers get to experience playing with and watching a proxy of him.

As a related scenario, whenever a mimetic model is available, there is the possibility that people will use it to have a "private audience" with a simulated version of the target. The age-old question "If you could have a conversation with any person, living or dead, who would it be?" may not be so hypothetical with mimetic models. Given access to the appropriate model, one could talk with a proxy of a famous world leader, a respected author, or a celebrity.

These scenarios raise the clear threat of targets being devalued, or even replaced, by their respective mimetic models. If interactors enjoy interacting with the model of A as much as—or more than—interacting with A themself, then A's position in social and economic marketplaces is compromised. In the more extreme versions of this scenario presented above, A's work could even be completely replaced by the work produced by A's model. It is important to note that this raises a new question for the future of work, as A's replacement is valuable because of A's unique qualities, which a mimetic model can capture but a traditional ML model cannot. In contrast, most of the discussion around automation and human labor has focused on situations in which humans performing generic tasks are replaced with generic machines. Here, individual people who currently have no substitutes at all, human or machine, are now threatened with the prospect of mimetic models that can partially or completely substitute for them. Chess champions such as Magnus Carlsen have traditionally commanded up to tens of thousands of dollars for the chance to play them in a single game. Similarly, top e-sports professionals are paid hefty appearance fees to participate in events. How might this change if mimetic versions of these players are widely available?

In addition to these labor considerations, people valued for unique traits, outputs, or interaction styles could find themselves devalued by the presence of mimetic models that capture their signature styles to a reasonable degree. If people are satisfied by having a private audience with mimetic proxies, the targets may consequently be less in demand. Individuals may lose some of their social capital if part of their uniqueness is lost to mimetic models. Perhaps even friends would be less in demand—if the mimetic version of your friend can do a convincing job of reacting to your stories or problems as they would, how will that affect your friendship?

Mimetic counterfactuals. Mimetic models, by generating realistic actions faithful to a specific individual's style, could be used to play out various counterfactual scenarios. For example, fans of creative endeavors often speculate what would have happened if person X had been in situation Y. For example, what would have happened if Bobby Fischer had shown up for his 1975 World Championship match with Anatoly Karpov instead of forfeiting it? How might Mozart's music have evolved if he had lived past 35? What did the letters that Nora Joyce wrote to her husband James Joyce contain before their grandson burned them? In principle, one could employ mimetic models to attempt to answer these kinds of questions. A Fischer model and Karpov model could face off under 1975-like conditions to shed light on who might have won; a Mozart model that can extrapolate from his earlier styles to his later styles could further extrapolate beyond his death; a Nora Joyce model could "respond" to James Joyce's still-existing letters (and we might even judge the Nora model's attempt to fill in the gaps by how faithfully a James Joyce model's response adheres to his actual reply).

Beyond historical questions, one could also explore contemporary counterfactuals via mimetic models. How would your idea for a song have turned out if you gave it to Taylor Swift? Which of your brilliant chess moves would the current World Champion Magnus Carlsen have failed to find? How might a debate between politicians go with the prompt you wish had been asked? Contemporary figures are just as easily modeled as historical ones, if not more easily due to the generally increased training data available.

In all of these scenarios, perhaps the most immediate ethical implication is the risk of reputational damage to the targets. To the extent that the models are imperfect representations of their targets, they will occasionally deviate from what the target would actually do. These deviations, especially salient or problematic ones, could alter what others think of the target. And more generally, we cannot know how accurate a mimetic model's behaviors are at extrapolating to a fully counterfactual scenario. If the mimetic simulation of the 1975 World Championship ends up with Karpov dethroning Fischer, that could alter the public's perception of these two players. If the Mozart model ends up reproducing musical innovations that others later conceived, the credit for them may shift. If the Nora Joyce model outputs offensive content, historians may think of her differently. We are familiar, for example, with similar effects arising from inaccurate public perception of real events based on historical fiction, such as when obituaries of Mark Felt (who served as Bernstein and Woodward's anonymous source in the Watergate scandal) attributed the quote "Follow the money" to him, despite the fact that this quote was uttered only by his fictional counterpart in William Goldman's screenplay for the movie *All the President's Men* [35].

Importantly, the risk of reputational damage in these counterfactual scenarios could actually *increase* with the accuracy of the models. If mimetic models aren't accurate, people will be less likely to trust them. A mimetic model that makes obvious or frequent mistakes would come across more as a caricature than a realistic representation. If one's expectations of the model are low, then mistakes, deviations, or questionable outputs could easily be attributed to quirks of the model rather than traits of the target. But if highly accurate mimetic models, such as those that already exist in chess and writing, were to generate the same mistakes or questionable outputs, they could be interpreted very differently. An accurate mimetic model engenders trust by generating realistic outputs, including ones we can validate by comparing with the target's actual response to the same input situations. Whatever outputs they generate will typically be treated as more reflective of the target rather than model artifacts.

In addition to the reputational damage that individual targets may suffer, mimetic models may be systematically biased in their misrepresentations. As a result, entire populations of people may be perceived worse because of how they are mischaracterized by mimetic models. Again, this risk is pronounced for generally accurate models that engender more trust by end-users. Although many ML models have been found to be systematically biased against particular subgroups, algorithmic bias that arises in mimetic models could pose new risks. Since mimetic models differ from each other by definition—as they target different individuals—systematic errors across a particular subgroup could be mistakenly attributed

to the subgroup rather than arising from correlated flaws across many different models.

2.3 Case Study: Mimetic Models of Oneself

An interesting case arises when considering the use of a mimetic model where the creator, operator, and target are the same person—in other words, when an individual creates and deploys a mimetic model of themself. Such a mimetic model may be used both as a means to an end and as an end in itself, traversing the scenarios discussed above and their associated ethical and social considerations.

Consider the use of a mimetic model of oneself as an end. One natural use case for such a model is as a stand-in for work: a mimetic model can perform work on a person's behalf without them having to expend any effort or time. For example, a person might create a mimetic model that predicts their own responses to messages [112], such as e-mail from work colleagues, and sends responses automatically on their behalf [45]. By creating and operating multiple models, the person can essentially use mimetic models as a *force multiplier*, to scale out their work and increase the number of people they interact with. For example, an artist specializing in portraiture could use a mimetic model to create portraits of customers, given a photograph, in a style that mimics what they would have created by hand [17, 107]. This would enable the artist to create many more custom portraits than would be physically possible.

In a similar vein, a mimetic model could enable someone to provide a private audience for multiple people at the same time. For instance, a sought-after chess coach could interact with multiple students at the same time by having them play against a mimetic model that captures the coach's playing style and decisions [64], providing each student with a private, one-on-one training experience. Although the coach could alternatively play an online simultaneous exhibition against the students, rotating through the games and making each move, this would be physically and mentally taxing for the coach, and the quality of each game would degrade as more students are added. In contrast, a mimetic model of the coach would not be subject to these physical limitations.

A mimetic model of oneself could also be used as a means to an end. One natural use case is to allow a mimetic model to interact with other people or entities before interacting with them in real life, as a way of filtering or preparing for these interactions. For example, a person who wishes to join an online dating site may be unfamiliar with the site's population or environment [84]. By creating a mimetic model of themself and allowing it to interact with the online site and its participants, they can observe the outcomes of these interactions and selectively pursue the interactions that seem most promising in real life.

While hypothetical, several of the above scenarios are within reach today. Large-scale language models have shown great promise in being fine-tuned to specific applications [14]; fine-tuning them to an individual's writing style is within reach. Personalized models of chess can already be trained with high enough accuracy to uniquely identify each player, given a moderate number of games per player [64]. And while artists, musicians, and authors have long used "ghost" assistants to scale out their work, the rise of mimetic models is bringing an unprecedented automation to this practice.

In all of these scenarios, the target, creator, and operator of the mimetic model are the same person. This presents a different subtlety to the ethical issues raised in previous sections, because issues of privacy or consent in the creation and use of the mimetic model diminish—the target of the model, being the individual themself, already embodies these rights—whereas issues of disclosure, value, and impact become more prominent. To start with, what level of disclosure is appropriate for the authorship of the mimetic model's communication and actions [93]? Should each e-mail message written by a mimetic model be explicitly flagged as such, so the recipient knows it was not written by a real person? What is the monetary value of artwork created by a mimetic model compared to artwork created by a real person? How do our answers change if the output produced by the mimetic model is perceived as better than what the target individual would have produced, or worse?

These questions underscore the importance of fidelity as a dimension for assessing the value of a mimetic model. If a mimetic model does a poor job of mimicking the target individual (i.e., it has low fidelity), then its value is clearly decreased and interactors will reject the model's similitude to the target. A more interesting situation arises when the mimetic model has high fidelity. In this case, even if interactions with the model faithfully simulate interactions with the real person, a person using multiple mimetic models of themself might potentially reduce the value of each interaction. Does a "thank you" e-mail sent by a mimetic model, however authentically crafted, evoke the same level of gratitude as a message written by the actual person? Should original artwork generated by a mimetic model command the same price as original artwork created by the actual person? Could a practice chess game with a mimetic model of a coach provide a better learning experience than a real game with the coach, if the coach is distracted or tired in real life? In all of these situations, the distinction between mimetic output and real output, and the relative quality of these outputs, influences the value that interactors will attribute to the corresponding interaction.

Note that the devaluation mentioned above may constitute an acceptable trade-off for a person: even if interactions with their mimetic models are valued less than interactions with the person in real life, the scalability of mimetic interactions could make them financially advantageous to the individual. For example, a chess coach might provide a discount for playing training games with their mimetic model (and support thousands of students simultaneously), while charging substantially more for playing with them in real life.

An interesting ethical consideration arises when a mimetic model of a person behaves differently than the person would, whether in a positive or negative sense. As a positive example, consider a mimetic model that responds to email using a level of politeness that is higher than the target individual's natural politeness. The responses may be adjusted to avoid language that some readers might find offensive; indeed, one can imagine a marketplace of apps that filter or modulate a mimetic model's output to achieve desirable properties. Such intentional modifications to a mimetic model's output could raise ethical considerations because they misrepresent the target individual and may be viewed as deceptive.

Mimetic models may also deviate from the target individual's behavior in a negative sense, for example by exaggerating a negative tendency However, since a mimetic model acts as a stand-in for

the target individual, its actions have direct implications for the individual's reputation and their liability in the event of harms being inflicted on the model's interactors. These harms extend beyond "noise" in the model training process and include endogenous biases that exist within the individual themself, which may be adopted or even amplified by the mimetic model. If the mimetic model is deployed at scale, this could result in the individual's biases being proliferated at scale. For example, if an individual who is prone to offensive comments creates and deploys mimetic models of themself on various online dating sites, this could amplify the effect of such individuals on these sites.

3 OVERVIEW OF ETHICAL THEMES

Having examined the ethical questions that arise when mimetic models are deployed in a range of specific scenarios, we now discuss some of the common themes that run through these scenarios, and their implications more generally.

Several themes recur in our analyses. First, the presence of mimetic models has the potential to significantly alter the relationships between people across a variety of settings. One of the simplest but clearest demonstrations of this is in competition: unequal access to mimetic models could substantially change the nature of who can compete, and the outcomes that can arise from competitions. This holds for both models used as a means to an end—e.g., in preparing for upcoming competitions—and as an end in themselves—e.g., in replacing real people with mimetic models of them. Relatedly, mimetic models have the potential to seriously change how individual people are valued. To the extent that individuals are valued in certain settings for their idiosyncratic behaviors and products, either socially, economically, or otherwise, and to the extent that mimetic models can faithfully simulate these behaviors, there may be significant effects on how people are valued. This also includes concepts of self-worth: how people value themselves could be influenced by how the interactions and outputs of their mimetic models are valued. An interesting consideration for a more distant future is how the value of human-ness itself might change in a world where mimetic models are powerful and commonplace. Will the role of friendship change if a mimetic model can fulfill some of the functions that human contact currently plays; or perhaps will in-person interactions with real people become more important, to guarantee that you are engaging with an actual person and not their mimetic model?

Another consistent theme across our scenarios is the increased capacity for bad-faith activities using mimetic models. Although we did not analyze deceptive practices in depth since they are already relatively common, mimetic models may make deception an even more prevalent threat. Imagine a phishing attack where a scammer pretends to be a trusted party, and can sustain a prolonged interaction posing as this trusted party. Mimetic models also increase the scope for manipulation. If one can thoroughly test how a particular target person will react to a wide variety of prompts or actions, it becomes more feasible to identify weaknesses that can be exploited for one's own benefit. Finally, the new privacy risks posed are easy to see. Mimetic models could qualitatively change our ability to process past behaviors and generalize to novel situations, thus raising the prospect of unintentionally leaking information about ourselves, our behaviors, and our identities.

Finally, we take note of three important dimensions of mimetic models that appear to play an influential role in determining the ethical consequences their use may have. First is the *fidelity* of the model, or how faithfully it captures its target's behaviors and characteristics. Many of the ethical issues we have discussed become more salient as model fidelity increases. If a mimetic model is only passably accurate, and is often easily distinguishable from the target, then it becomes more of a caricature than a realistic simulation. As such, issues such as deceptive practices and reputational damage become less of a concern. Second is the *modality* of the model, the domain it operates in and the types of behaviors it is designed to reflect. Clearly, a model that can output text differs from one that can output chess moves, and the ethical issues raised by each differs as a result. Third is the *generality* of the model, or the breadth of scenarios and domains that a mimetic model can capture. Generally speaking, the wider the model's reach, the more pertinent the ethical concerns.

4 RELATED WORK

4.1 General Considerations

Some of the initial discussions of mimetic models occurred in science fiction (e.g. [10, 26, 76, 103]), but our understanding of them has become much more specific as the technology to produce them has become concrete and increasingly available. Our discussion of the normative considerations related to mimetic models in turn connects to some of the central themes in the ethics of AI, including the fairness of decisions [7, 16, 29, 90], the potential for bias [12, 39, 50, 80], and potential shifts in accountability [49].

Mimetic models also introduce questions related to data access [114] and informed consent [22], and may benefit from strategies such as *Model Cards* [67] to address these issues. When mimetic models are produced on anonymized data, they introduce the risk of deanonymization through their behavior, based on some of the principles in the privacy literature [72, 113]. Mimetic models contain significant potential for deception as well, and the issues here are related to the issues that arise with *deepfakes* [102], as we discuss next. Some of the concerns associated with this type of deception are fake announcements by public figures [3], devaluing of performers [88], and fake news [111].

4.2 Related Concepts

As noted in the introduction, it is useful to explore the relationship between mimetic models and related concepts at the boundary of AI modeling and human behavior. We consider a number of these in this subsection.

4.2.1 Deepfakes. Deepfakes raise normative concerns that overlap those encountered with mimetic models. The term *deepfake* refers to a set of techniques for manipulating video or images to replace or generate the likeness of a person.[†] The name originates from a deep-learning face-swapping program, popularized by the Reddit user *deepfakes*, that allows a user to replace the face of an actor

[†]The techniques are not limited to humans, but we focus on their application to humans here.

in a video (or still image) with that of another target [99, 101]. Importantly, the requirements for training the model are low, the system can be run on a single consumer-level GPU, the replaced video can be low resolution, and the number of samples required for the target can be as little as a single image.‡

Expanding beyond this specific origin, the term *deepfake* has grown to acquire a broader definition in the culture more generally (e.g. [4]), and is now viewed as a key component in *fake news* [54, 56]. As noted in the introduction, a key distinction between even this broader framing of deepfakes and the concept of a mimetic model is the fact that mimetic models are designed for interaction in new situations. We require mimetic models to be able to interact with people, in which they take some action, observe the response, and take another action based on the response. In contrast, deepfakes are typically pre-generated for a single planned behavior. One way to think of the relationship is to note that a mimetic model could naturally be used to generate the text spoken by a deepfake model. Of course, the distinction is not absolute, and adding interactivity to a deepfake would produce a type of mimetic model.

4.2.2 Digital Avatars. Many people employ visualization of their online persona that is distinct from their own physical body, be it a simple cartoon image or a complex 3D model [24, 95]. These avatars act on behalf of the "target", to use our framework's terminology, either directly under the control of the target or in some pre-programmed way. Thus the concerns that misuse or mistreatment [21, 40] of avatars raise has overlap with those of mimetic models. Additionally, people can become attached to their avatars both emotionally [31, 105] and through their physical representation [75, 108]; having a virtual representation of yourself can in some cases lead to a phenomenon known as the *Proteus effect* [109, 110], in which people adapt their behavior based on characteristics of the avatar. The use of avatars to test new experiences overlaps with the use of mimetic models as proxies, as we discuss in Section 2.2.

4.2.3 Style Transfer. Style transfer [32, 55, 57] is a technique in which an algorithm transforms a piece of media to render it in the style of a specified target author. Style transfer techniques typically use a single static initial image [32], video [89], audio clip [18], or other representation [60]. However, at a broader level of abstraction, they can be viewed as creating a special-purpose mimetic model of the target author, for the purpose of interacting with a prompt to produce new work in the target author's style.

4.2.4 Multi-modal generative agents. There has been ongoing progress in machine learning systems that translate prompts such as 'A Mayan warrior getting ready, in the style of Rembrandt' [77] into an image matching the prompt—e.g., ImageBERT [78], ALIGN [47], CLIP [79] and DALL·E 2 [82, 83], or the reverse (images to text) like Flamingo [5]. These systems allow for outputs that mimic the styles of specific individuals, and can be fine-tuned to allow for style transfer [59]. Generating mimetic models is not the main goal of these works, but they may be the foundation for mimetic models.

‡More angles/lighting conditions lead to a better result, so multiple images are required to generate a more dynamic set of outputs.

4.2.5 Model Personalization. Personalized systems are those that adapt their outputs to the user they are interacting with [61]. This is often done by creating a model that interacts with a user over time, maintaining and improving a representation of the model's knowledge about the person [48, 106]. The task is thus a type of mirror image to what a mimetic model does: personalization seeks to create a model that can make optimal responses to a user, while a mimetic model instead seeks to act as a stand-in for the user and generate responses in their stead.

4.2.6 Legal Stand Ins. A non-computational analogy to mimetic models in the off-line world can be found in the way that legal systems allow for proxies [25], power-of-attorney, or other mechanisms to allow a designated individual to make decisions that are intended to represent the intent of a specific target person. As a result, the history of ethical considerations involving proxies can provide insights into the corresponding issues that may arise with mimetic models [87].

4.2.7 Other Concepts. Finally, we touch on a few additional concepts more briefly.

Recommender Systems. Systems that recommend content by modeling a user's preferences [9, 85] are not mimetic in our sense, since they are not generating behavior on behalf of the user. However, we can imagine ways in which mimetic modeling ideas could be incorporated into a larger recommendation context, such as through mimetic modeling of the next movie selected to play (i.e. autoplay behavior) [46].

Work Automation. There is of course a vast literature on automation, and the ways in which AI in particular is replacing certain categories of jobs [1]. Our analysis overlaps with this literature only to the extent that jobs are being replaced by models of *specific* workers, rather than the typical practice of designing AI or ML systems to perform well on the underlying task in a generic or aggregate sense. This distinction also applies in the context of automation via robotics [2, 51].

Prediction. There are well-established methodologies for converting a generative system to a predictive one [73], and via this principle mimetic models can be used to predict a person's behavior, simply by observing the behavior that is generated by the model. This translation implies that mimetic models share the same concerns about predicting the behavior of individuals [20].

Speculative Fiction. As noted at the start of this section, many of the ethical issues we discuss here are also found in works of fiction [27, 62]. Fictional approaches to these questions are not bounded by real-world constraints, and so they are often much more exaggerated in their formulations than what we consider here. For example, works like David Brin's *Kiln People* [10], Greg Egan's *Zendegi* [26] or Vernor Vinge's *The Cookie Monster* [103] all directly discuss the implications of high-fidelity models of specific people and their ethical implications.

4.3 Current Applications of Mimetic Models

One of the main realized uses of mimetic models in practice today is for game-playing. As a general domain, games provide both highly detailed behavioral data [41, 66] and easy creation of computer-controlled players [11]. There are also financial incentives for game

vendors to provide mimetic models as a feature for users [34, 71, 97, 98]. Chess [63], Go [91] and other [43] tabletop games have also been studied in the context of creating human-like models, usually with a focus on human-compatible agents [42] or creating tools for teaching humans [64].

Mimetic models have also been investigated in educational settings, with the creation of models of both student [33] and teacher [13] behavior. In these cases, however, the generative nature of the models was not the focus of the research. Content-filling algorithms such as in-painting brushes [38, 52] can also be viewed as a type of mimetic model, raising similar issues to applications in text generation discussed in Section 2.3. Finally, mimetic models have been used to encode individual artistic style; one example is in archaeological studies of pottery [81], where the goal is to generate similar pieces of pottery based on the styles of specific artisans, or models like DALL·E 2 [82] that can create an image in the style of a specific artist matching a prompt.

5 CONCLUSION

Mimetic models represent a complex new direction in the use of AI to model human behavior—one in which models are tailored to match the behavior of specific individuals, and in settings that allow for rich interaction with others. We have seen that mimetic models surface subtle ethical and social considerations across a wide range of scenarios—including as forms of preparation for future interactions with real people (in a competition, an interview, or a date); as an end in themselves to study counterfactuals or to provide spectator experiences that would be hard to produce using real people; and as a way for people to create realistic stand-ins for themselves. We believe that the framework here suggests a number of directions for further investigation, including more extensive domain-specific considerations as more powerful mimetic models become available across an increasingly wide array of contexts.

REFERENCES

[1] Acemoglu, D., and Restrepo, P. Automation and new tasks: How technology displaces and reinstates labor. *Journal of Economic Perspectives 33*, 2 (May 2019), 3–30.

[2] Acemoglu, D., and Restrepo, P. Robots and jobs: Evidence from us labor markets. *Journal of Political Economy 128*, 6 (2020), 2188–2244.

[3] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., and Li, H. Protecting world leaders against deep fakes. In *CVPR workshops* (2019), vol. 1.

[4] Ajder, H., Patrini, G., Cavalli, F., and Cullen, L. The state of deepfakes landscape, threats, and impact, 2019.

[5] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning. *ArXiv abs/2204.14198* (2022).

[6] Bard, N., Foerster, J. N., Chandar, S., Burch, N., Lanctot, M., Song, H. F., Parisotto, E., Dumoulin, V., Moitra, S., Hughes, E., et al. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence 280* (2020), 103216.

[7] Barocas, S., Hardt, M., and Narayanan, A. Fairness in machine learning. *NeurIPS tutorial 1* (2017), 2.

[8] Bird, J. J., Faria, D. R., Ekárt, A., Premebida, C., and Ayrosa, P. P. Lstm and gpt-2 synthetic speech transfer learning for speaker recognition to overcome data scarcity. *arXiv preprint arXiv:2007.00659* (2020).

[9] Bobadilla, J., Ortega, F., Hernando, A., and Gutiérrez, A. Recommender systems survey. *Knowledge-based systems 46* (2013), 109–132.

[10] Brin, D. *Kiln People.* Tor, 2002.

[11] Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540* (2016).

[12] Buolamwini, J., and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (2018), PMLR, pp. 77–91.

[13] Chaturvedi, S., Goldwasser, D., and Daumé III, H. Predicting instructor's intervention in mooc forums. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2014), pp. 1501–1511.

[14] Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., and Zaremba, W. Evaluating large language models trained on code. *CoRR abs/2107.03374* (2021).

[15] Chopra, S., Gianforte, R., and Sholar, J. Meet percy: The cs 221 teaching assistant chatbot. *ACM Transactions on Graphics 1*, 1 (2016), 1–8.

[16] Chouldechova, A., and Roth, A. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM 63*, 5 (2020), 82–89.

[17] Ci, Y., Ma, X., Wang, Z., Li, H., and Luo, Z. User-guided deep anime line art colorization with conditional adversarial networks. In *Proceedings of the 26th ACM international conference on Multimedia* (2018), pp. 1536–1544.

[18] Cífka, O., Şimşekli, U., and Richard, G. Groove2groove: One-shot music style transfer with supervision from synthetic data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing 28* (2020), 2638–2650.

[19] Cimini, A. Walking to the gallery: Sondra Perry's "It's in the game" in San Diego in five fragments. *Sound Studies 4*, 2 (2018), 178–200.

[20] Crawford, K., and Schultz, J. Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev. 55* (2014), 93.

[21] Dechant, M. J., Birk, M. V., Shiban, Y., Schnell, K., and Mandryk, R. L. How avatar customization affects fear in a game-based digital exposure task for social anxiety. *Proceedings of the ACM on Human-Computer Interaction 5* (2021), 1 – 27.

[22] Department of Health, Education, and Welfare, and National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The belmont report. ethical principles and guidelines for the protection of human subjects of research. *The Journal of the American College of Dentists 81*, 3 (2014), 4–13.

[23] Dhou, K. Towards a better understanding of chess players' personalities: A study using virtual chess players. In *International Conference on Human-Computer Interaction* (2018), Springer, pp. 435–446.

[24] Ducheneaut, N., Wen, M.-H., Yee, N., and Wadley, G. Body and mind: a study of avatar personalization in three virtual worlds. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2009).

[25] Easterbrook, F. H., and Fischel, D. R. Voting in corporate law. *The journal of Law and Economics 26*, 2 (1983), 395–427.

[26] Egan, G. *Diaspora.* Gollancz, 2002.

[27] Egan, G. *Zendegi.* Gollancz, 2010.

[28] Fazelpour, S., and Lipton, Z. C. Algorithmic fairness from a non-ideal perspective. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2020), pp. 57–63.

[29] Finocchiaro, J., Maio, R., Monachou, F., Patro, G. K., Raghavan, M., Stoica, A.-A., and Tsirtsis, S. Bridging machine learning and mechanism design towards algorithmic fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021), pp. 489–503.

[30] Fish, B., and Stark, L. Reflexive design for fairness and other human values in formal models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (2021), pp. 89–99.

[31] Fox, J., and Bailenson, J. N. Virtual self-modeling: The effects of vicarious reinforcement and identification on exercise behaviors. *Media Psychology 12* (2009), 1 – 25.

[32] Gatys, L. A., Ecker, A. S., and Bethge, M. Image Style Transfer Using Convolutional Neural Networks, 2016.

[33] Geigle, C., and Zhai, C. Modeling MOOC Student Behavior With Two-Layer Hidden Markov Models. In *L@S '17: Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale.* Association for Computing Machinery, New York, NY, USA, Apr 2017, pp. 205–208.

[34] Gitlin, J. War Stories: How Forza learned to love neural nets to train AI drivers, Dec 2021. [Online; accessed 7. Dec. 2021].

[35] Greenberg, D. William goldman: The writer who brought watergate to the screen. *Politico* (December 2018).

[36] Guan, M., Gulshan, V., Dai, A., and Hinton, G. Who said what: Modeling individual labelers improves classification. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2018), vol. 32.

[37] Guo, W., and Caliskan, A. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (2021), pp. 122–133.

[38] Guzdial, M., Liao, N., Chen, J., Chen, S.-Y., Shah, S., Shah, V., Reno, J., Smith, G., and Riedl, M. O. Friend, collaborator, student, manager: How design of an ai-driven game level editor affects creators. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (2019), pp. 1–13.

[39] Harcourt, B. E. Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter 27*, 4 (2015), 237–243.

[40] Hill, D. W. Avatar ethics: Beyond images and signs. *Journal for Cultural Research 17* (2013), 69 – 84.

[41] Hooshyar, D., Yousefi, M., and Lim, H. Data-driven approaches to game player modeling: a systematic literature review. *ACM Computing Surveys (CSUR) 50*, 6 (2018), 1–19.

[42] Hu, H., Lerer, A., Cui, B., Pineda, L., Brown, N., and Foerster, J. Off-belief learning. In *International Conference on Machine Learning* (2021), PMLR, pp. 4369–4379.

[43] Jacob, A. P., Wu, D. J., Farina, G., Lerer, A., Bakhtin, A., Andreas, J., and Brown, N. Modeling strong and human-like gameplay with kl-regularized search. *arXiv preprint arXiv:2112.07544* (2021).

[44] Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., Castaneda, A. G., Beattie, C., Rabinowitz, N. C., Morcos, A. S., Ruderman, A., et al. Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science 364*, 6443 (2019), 859–865.

[45] Jahanshahi, H., Kazmi, S., and Cevik, M. Auto response generation in online medical chat services. *arXiv preprint arXiv:2104.12755* (2021).

[46] Jenner, M. Is this tviv? on netflix, tviii and binge-watching. *New media & society 18*, 2 (2016), 257–273.

[47] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML* (2021).

[48] Kang, W.-C., and McAuley, J. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)* (2018), IEEE, pp. 197–206.

[49] Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. Human decisions and machine predictions. *The quarterly journal of economics 133*, 1 (2018), 237–293.

[50] Kleinberg, J., Ludwig, J., Mullainathan, S., and Sunstein, C. R. Discrimination in the age of algorithms. *Journal of Legal Analysis 10* (2018), 113–174.

[51] Kober, J., Bagnell, J. A., and Peters, J. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research 32*, 11 (2013), 1238–1274.

[52] Koch, J., Lucero, A., Hegemann, L., and Oulasvirta, A. May ai?: Design ideation with cooperative contextual bandits. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019).

[53] Kokkinakis, A., York, P., Patra, M., Robertson, J., Kirman, B., Coates, A., Pedrassoli Chitayat, A., Demediuk, S. P., Drachen, A., Hook, J. D., et al. Metagaming and metagames in esports. *International Journal of Esports* (2021).

[54] Korshunov, P., and Marcel, S. *The Threat of Deepfakes to Computer and Human Visions*. Springer International Publishing, Cham, 2022, pp. 97–115.

[55] Krishnan, P., Kovvuri, R., Pang, G., Vassilev, B., and Hassner, T. TextStyleBrush: Transfer of Text Aesthetics from a Single Example. *arXiv* (Jun 2021).

[56] Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., et al. The science of fake news. *Science 359*, 6380 (2018), 1094–1096.

[57] Li, Y., Wang, N., Liu, J., and Hou, X. Demystifying Neural Style Transfer. *arXiv* (Jan 2017).

[58] Liang, C., Proft, J., Andersen, E., and Knepper, R. A. Implicit communication of actionable information in human-ai teams. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–13.

[59] Liu, Z.-S., Wang, L.-W., Siu, W. C., and Kalogeiton, V. S. Name your style: An arbitrary artist-aware image style transfer. *ArXiv abs/2202.13562* (2022).

[60] Ma, C., Ji, Z., and Gao, M. Neural style transfer improves 3d cardiovascular mr image segmentation on inconsistent data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019), Springer, pp. 128–136.

[61] McAuley, J. *Personalized Machine Learning*. Cambridge University Press, 2022.

[62] McCarthy, W. *The wellstone*. Bantam Books, 2003.

[63] McIlroy-Young, R., Sen, S., Kleinberg, J., and Anderson, A. Aligning superhuman ai with human behavior: Chess as a model system. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2020), pp. 1677–1687.

[64] McIlroy-Young, R., Wang, R., Sen, S., Kleinberg, J., and Anderson, A. Learning models of individual human behavior in chess. In *Proceedings of the 28th ACM SIGKDD international conference on Knowledge discovery and data mining* (2022).

[65] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR) 54*, 6 (2021), 1–35.

[66] Melhart, D., Azadvar, A., Canossa, A., Liapis, A., and Yannakakis, G. N. Your Gameplay Says It All: Modelling Motivation in Tom Clancy's The Division. In *2019 IEEE Conference on Games (CoG)*. IEEE, 2019, pp. 20–23.

[67] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (2019), pp. 220–229.

[68] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature 518*, 7540 (2015), 529–533.

[69] Moravčík, M., Schmid, M., Burch, N., Lisỳ, V., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M., and Bowling, M. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science 356*, 6337 (2017), 508–513.

[70] Morgenstern, O., and Von Neumann, J. *Theory of games and economic behavior*. Princeton university press, 1953.

[71] Morris, C. Former NCAA athletes win video game lawsuit against EA. *NBC News* (Aug 2013).

[72] Narayanan, A., and Shmatikov, V. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)* (2008), IEEE, pp. 111–125.

[73] Ng, A., and Jordan, M. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems 14* (2001).

[74] Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. Gaugan: Semantic image synthesis with spatially adaptive normalization. In *ACM SIGGRAPH 2019 Real-Time Live!* (New York, NY, USA, 2019), SIGGRAPH '19, Association for Computing Machinery.

[75] Peck, T. C., Seinfeld, S., Aglioti, S. M., and Slater, M. Putting yourself in the skin of a black avatar reduces implicit racial bias. *Consciousness and Cognition 22* (2013), 779–787.

[76] Pohl, F. *Gateway*. St. Martin's Press, 1977.

[77] Porres, D. A mayan warrior getting ready, in the style of rembrandt. https://twitter.com/PDillis/status/1530297800453496833, May 2022.

[78] Qi, D., Su, L., Song, J., Cui, E., Bharti, T., and Sacheti, A. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *ArXiv abs/2001.07966* (2020).

[79] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (2021), PMLR, pp. 8748–8763.

[80] Raghavan, M., Barocas, S., Kleinberg, J., and Levy, K. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (2020), pp. 469–481.

[81] Ramazzotti, M., Buscema, P. M., Massini, G., and Della Torre, F. Encoding and simulating the past. a machine learning approach to the archaeological information. In *2018 Metrology for Archaeology and Cultural Heritage (MetroArchaeo)* (2018), IEEE, pp. 39–44.

[82] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

[83] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. *ArXiv abs/2102.12092* (2021).

[84] Ranzini, G., and Lutz, C. Love at first swipe? explaining tinder self-presentation and motives. *Mobile Media & Communication 5*, 1 (2017), 80–101.

[85] Resnick, P., and Varian, H. R. Recommender systems. *Communications of the ACM 40*, 3 (1997), 56–58.

[86] Ressmeyer, R., Masling, S., and Liao, M. "deep faking" political twitter using transfe r learning and gpt-2, 2019.

[87] Rezaee, Z. *Corporate governance and ethics*. John Wiley & Sons, 2008.

[88] Rosner, H. The Ethics of a Deepfake Anthony Bourdain Voice in "Roadrunner". *New Yorker* (Jul 2021).

[89] Sanakoyeu, A., Kotovenko, D., Lang, S., and Ommer, B. A style-aware content loss for real-time hd style transfer. In *proceedings of the European conference on computer vision (ECCV)* (2018), pp. 698–714.

[90] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency* (2019), pp. 59–68.

[91] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature 529*, 7587 (2016), 484–489.

[92] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science 362*, 6419 (2018), 1140–1144.

[93] Simmons, M., and Lee, J. S. Catfishing: A look into online dating and impersonation. In *International Conference on Human-Computer Interaction* (2020), Springer, pp. 349–358.
[94] Stein, R. What We Learned—Solving Standard - Hipsters of the Coast, Nov 2015.
[95] Suh, K.-S., Kim, H., and Suh, E.-K. What if your avatar looks like you? dual-congruity perspectives for avatar use. *MIS Q. 35* (2011), 711–729.
[96] Taddeo, M., and Floridi, L. How ai can be a force for good. *Science 361*, 6404 (2018), 751–752.
[97] Tantaros, A. Electronic Arts, identity thief? *Nydailynews* (Jan 2019).
[98] Thomas, K. Sports Video Game Suit Gets to Heart of First Amendment Clash. *N.Y. Times* (Nov 2010).
[99] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., and Ortega-Garcia, J. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion 64* (2020), 131–148.
[100] Tomašev, N., Paquet, U., Hassabis, D., and Kramnik, V. Assessing game balance with alphazero: Exploring alternative rule sets in chess. *arXiv preprint arXiv:2009.04374* (2020).
[101] Tora, M. Faceswap, 2018.
[102] Vaccari, C., and Chadwick, A. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society 6*, 1 (2020), 2056305120903408.
[103] Vinge, V. *The Cookie Monster.* Analog Science Fiction and Fact, 2003.
[104] Whittaker, M., Alper, M., Bennett, C. L., Hendren, S., Kaziunas, L., Mills, M., Morris, M. R., Rankin, J., Rogers, E., Salas, M., et al. Disability, bias, and ai. *AI Now Institute* (2019).
[105] Wolfendale, J. My avatar, my self: Virtual harm and attachment. *Ethics and Information Technology 9* (2006), 111–119.
[106] Wu, S., Tang, Y., Zhu, Y., Wang, L., Xie, X., and Tan, T. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence* (2019), vol. 33, pp. 346–353.
[107] Xu, P., Hospedales, T. M., Yin, Q., Song, Y.-Z., Xiang, T., and Wang, L. Deep learning for free-hand sketch: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
[108] Yee, N., and Bailenson, J. N. The proteus effect: The effect of transformed self-representation on behavior. *Human Communication Research 33* (2007), 271–290.
[109] Yee, N., Bailenson, J. N., and Ducheneaut, N. Implications of transformed digital self-representation on online and offline behavior. *Communication Research 36* (2009), 285–312.
[110] Yee, N., Bailenson, J. N., and Ducheneaut, N. The proteus effect. *Communication Research 36* (2009), 285 – 312.
[111] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. Defending against neural fake news. *Advances in neural information processing systems 32* (2019).
[112] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. Personalizing Dialogue Agents: I have a dog, do you have pets too? *arXiv* (Jan 2018).
[113] Zimmer, M. "but the data is already public": on the ethics of research in facebook. In *The Ethics of Information Technologies*. Routledge, 2020, pp. 229–241.
[114] Zwitter, A. Big data ethics. *Big Data & Society 1*, 2 (2014), 2053951714559253.