Ethics of Artificial Intelligence:
From Dating to Finance

Edited by

Walter Sinnott-Armstrong
Kenzie Doyle

2019

TABLE OF CONTENTS

PREFACE

Walter Sinnott-Armstrong and Kenzie Doyle, Editors

Ethics 490 serves as the Capstone course for the Ethics Certificate Program at the Kenan Institute for Ethics at Duke University. The course comprises students from a variety of majors and minors who have completed an extensive array of ethics courses required for the Certificate. In Ethics 490 the students bring their diverse perspectives and expertise to bear on a contemporary moral issue. The Spring 2018 iteration of Ethics 490 focused on ethical issues raised by artificial intelligence.

The goal, as in previous installments of the course, was to write a book together. Starting in January, we spent the first half of the term locating, reading, and discussing background and recent materials and began to narrow our attention to specific issues in the ethics of artificial intelligence. Then, in the second half of the term, each student chose a particular application of artificial intelligence and wrote several drafts of a chapter on that topic. Students also read and commented on each other's draft chapters. We discussed every chapter together in class. By the end of this process, each student had a polished piece of work to be proud of. The result is this book.

The essays collected here address many kinds of artificial intelligence as well as a variety of ethical issues raised by artificial intelligence. While artificial intelligence continues to bring new benefits as well as dangers, these essays chart exciting progress in our understanding of these important phenomena.

We want to thank the students for their hard work, for helping each other, for believing in themselves, and for inspiring us. Together you have produced an impressive volume. Thanks also to Suzanne Shanahan and the Kenan Institute for Ethics for supporting this project. The editors are also grateful for a grant from the Templeton World Charity Foundation that facilitated this work. Of course, the contents of this book are solely the responsibility of the authors and do not necessarily represent the official views of the Templeton World Charity Foundation, the Kenan Institute for Ethics, or Duke University.

INTRODUCTION

Kenzie Doyle

Artificial Intelligence (AI) is present in many aspects of our daily life and is becoming more ubiquitous. Autonomous agents interact with humans in increasingly sophisticated ways, taking important roles in a wide range of fields from transportation to medicine and politics. These autonomous agents provide services with great potential to benefit society, but they also create significant ethical concerns. This book aims to examine the moral considerations relevant to some of the real uses of AI in 2019.

What Is AI?

Definitions of what is or is not AI change as technology advances. For the purpose of this volume, an AI is defined as a non-organic agent capable of learning in order to generate original content that humans cannot anticipate. The AI might produce its results by using massive quantities of data to analyze trends, by performing calculations too complex for human statisticians, or by executing algorithmic operations. However, big data, computations, and algorithms are not enough to count as AI. Technology that is able to process large quantities of data is invaluable to humans, but many programs perform massive calculations that are dictated by humans without changing their own code in response to new information. In contrast, AI must be capable of reprogramming itself, rather than merely following the strict instructions of its human developer. Its results cannot be predicted by the humans who programmed it: the conclusions at which it arrives are the outcomes of new data processing strategies that the agent itself developed to adapt to new input. In short, AI can learn.

Central Ethical Issues

Applications of AI are diverse, but they share many ethical issues. The most prominent of those ethical issues are bias, equality, access, privacy, deception, autonomy, and freedom, as well as what it means to be human. Each of these will be described briefly below.

Bias: A significant component of many human endeavors is reliance on heuristics. Common heuristics can help us function in a complex social world, but they also lead to undesirable prejudice against disadvantaged groups that AI can help to counteract. Some AI is built with the intention of removing bias, intentional or otherwise, from decision making procedures so as to reduce discrimination. However, because AI is trained using human data from an unequal world, its behavior might exacerbate existing inequalities between groups. In this way, AI might incorporate human bias into its own procedures without "realizing" it, and without any innate prejudice of its own.

Equality: AI can help to resolve issues of inequality by providing new opportunities to disadvantaged groups and individuals, but some other uses of AI make current inequity worse. The way AI is deployed and distributed might benefit some people more than others, thus widening existing gaps based on income, race, geographic region, or any number of other factors. Additional problems arise when AI replaces jobs previously held by humans, especially

in fields that pay workers relatively low wages, but does not contribute negatively to unemployment in higher income fields.

Access: Another way in which AI might have an impact on equality is through access. Some uses of AI make resources or services more readily available, so they become accessible to a larger proportion of the population. However, in other cases, the technology might be expensive or proprietary, such that it only benefits certain, often already privileged, groups. This happens, for example, when large investors have access to AI that makes their stocks increase in value at a faster rate than those of smaller investors.

Privacy: AI relies on information, so questions of privacy arise when considering what information AI can be given and how it may use that information. Many people do not know what is being done with their data when they begin using a service. They also do not know whether or not their data can be easily traced back to them. AI can be used to identify people through use of big data, facial recognition, or other means, and how and when it is permissible to use AI to identify individuals is an important ethical issue. Even when data is given with full informed consent, privacy breaches can happen if the technology is hacked or misused. The security of the AI therefore has important ethical implications as well.

Deception: AI has the potential to simulate reality very effectively, to a degree that many humans cannot tell when something was generated by a machine. This is obviously troubling in cases in which, for example, a computer creates a "deep fake" video in which a person appears to be doing something that they never really did. Accusations of deception might also be valid in subtler instances, such as when providers of a service neglect to disclose that they are using AI.

Autonomy: One of the main benefits of AI is its ability to behave autonomously, such that humans need not provide constant oversight and can therefore spend their time elsewhere. However, it is not always clear just how much autonomy machines should be given. Sometimes questions arise from the amount of training the AI must undergo before being allowed to function with less human supervision, but sometimes the issues come more from the danger of human reliance on AI to the point that they no longer understand how decisions are being made. In certain cases, whether or not an AI can have the necessary capacities to make "good" decisions is unclear, such as in military contexts, when a machine that cannot feel emotions might not be deemed qualified to make choices related to the value of human life.

Freedom: Some uses of AI might benefit humanity in some ways, but negatively impact one's ability to act freely. This issue can be related to privacy concerns, if AI surveillance decreases anonymity. AI can also hinder free choice, such as when algorithms guide people to make certain decisions through selective advertising.

The nature of humanity: Even in the absence of other ethical concerns, some people might fear AI if they perceive that it destabilizes what it means to be human. If AI can create art, express emotions, communicate meaningfully, make moral decisions, and become the object of genuine human attachment, what makes humanity unique? Will AI always be a glorified computer program that is only an inferior substitute for humans in some ways, or can it become something more like us? Is there ever a point at which AI deserves personhood culturally, legally, and morally? And is that question's answer, whatever it might be, a good thing, or a bad thing?

All of these issues arise in the chapters to follow.

In This Book

As the title implies, this collection explores the ethical implications of a wide range of AI applications. Some uses of AI are commonplace and fun, such as in dating culture. In "AI in Love," Sloan Talbot investigates the nuances of personal preference relative to systematic bias in dating platforms, and identifies questions about how non-human agents should be allowed to influence romance. Applications of AI that people engage with regularly in daily life are not always used in optional contexts, and can have serious implications. Facial recognition algorithms use AI to identify faces in images, which can be applied for a variety of purposes from crime monitoring to unlocking smartphones, but which also could cause problems related to discrimination and privacy, as John Benhart describes in "Assessing Facial Recognition: An Ethical Exploration." AI can also help to combat some of the very same issues, too, as Onuoha Odim demonstrates in "AI and an Algorithmic Approach to Combat Gerrymandering," an investigation of how algorithmic approaches to redistricting can best be implemented to avoid biased agendas.

Whether AI exacerbates or helps to resolve ethical problems depends upon whether it is adequately regulated and transparent. Analese Bridges focuses on the legal aspect of systemic use of AI that other chapters identify: "I spy with my little eye? — Artificial Intelligence and Fourth Amendment Search" discusses the role of artificial intelligence in governmental searches by reviewing the US Supreme Court case Kyllo v. United States (2001) on thermal imaging technology, which found that technology is not in public use, but which might be challenged in the future as the use of AI continues to rise. In "The Ethics of AI in Journalism," Jackie Park reviews current use of AI in journalism, analyzing how AI interacts with the media's ethical obligations in democratic societies and pinpointing factors that need further consideration going forward as AI is implemented in newsrooms. The government and the news are both crucial elements of AI's ethical usage and in its further adoption based on public trust of it.

It is vital that AI be trustworthy: as technology develops, it has an increasingly ubiquitous presence in people's personal decision-making, as well as in the economic underpinnings of society. In "Ethical Concerns of Targeted Advertising," Jillian Kohn focuses on AI-driven advertising strategies that allow companies to reach their target audiences with great precision, but that also create complications related to discrimination and transparency because of the necessary use of private data. In "Applications of AI in the Financial Services Sector," Coleman Kraemer illustrates that the ethical implications of AI can vary widely depending on its specific use by analyzing the role of AI in robo-advising platforms and hedge fund trading strategies: two unique examples of how AI can both benefit less sophisticated investors and how it can aid more advanced investment managers in predicting market prices.

That AI now impacts the world in such far-reaching ways means also that it impacts people at a very personal, private level as well: AI is increasingly involved in healthcare. In "Artificial Intelligence and Healthcare," Shweta Lodha discusses the various ways AI is transforming the healthcare industry more broadly, and how the use of AI in medicine impacts patient privacy, informed consent, and propagation of medical bias. Furthermore, AI is increasingly being used in mental healthcare, as Sheridan Wilbur explores in "Is Virtual Reality a Viable Form of Treatment for Anxiety Disorders?" by considering virtual reality as a form of psychological therapy for anxiety disorders: virtual reality can make exposure-based therapy in a controlled setting more immersive than traditional methods, but it can also be a source of concern regarding accessibility, equality, and deception, and threats to jobs.

In the nine chapters in this volume, the authors contextualize the role of AI in the year 2019 and anticipate how its role will change going forward. The book aims to provide a nuanced and wide-ranging overview of the importance of AI with the understanding that, as technology advances, so too must ethics advance to guide it.

AI in Love

Sloan Talbot

There's a popular phrase in the U.S. about dating and love that goes, "There's plenty of fish in the sea."[1] The phrase is used usually in encouragement after a breakup to tell someone not to worry because there are many other people in the world who are single and eligible to date. While this phrase may come across as comforting to some, the reality is *there are* a lot of people in the world who are eligible to date; in fact, there might even be too many to choose from. Luckily, in the last decade, the tech industry has made meeting potential matches easier than ever before with the invention of dating websites and online dating apps. Seemingly perfect for a user who doesn't want to be overwhelmed with a lot of people, the dating site or app gives you a range of people to select to "match" with by applying sorting filters like distance away, age, and potentially many more specific preferences. Yet, these new conveniences come with questions and concerns.

Recently, the implementation of Artificial Intelligence, (AI) in dating sites has given way to further curation of a select dating pool tailored to the individual. Some have gone as far as to give potential first-date conversation starters, pick-up lines, and with your consent, calculate the expected longevity of this match based on personality traits. AI in dating can help to solve the problem of "too many fish in the sea" for someone trying to find a potential match, but it also brings with it ethical questions that need to be examined. With the new implementation of AI, we have to consider *how* the AI are selecting the profiles we get to view and see. Are there implicit biases from society that the AI is continuing to employ like European standards of beauty, and certain favorable socio-economic statuses? The AI has the potential to base the profiles it shows to the user on the ads they look at, the social media accounts they follow, and even the socio-economic status (SES) of the zip-code they live in or the race they identify with. Our society has demonstrated its readiness to adopt technology such that it is not the question of *if* this should be implemented, but *how* it is implemented. Moral and ethical implications of race and class bias in dating apps and sites, the future implications of having partners "selected" for us by a machine are questions and concerns that arise with this new wave of online romance.

We all have biases and preferences when it comes to whom we find attractive, so would an AI that scans and filters potential dates be that morally wrong if it is simply using the biases and filters we already implement? This chapter will first consider the role of preference and bias in romance and love, then delve into how AI in apps and sites can either help us to navigate our preferences, or heighten biases already rampant in our society. It will also distinguish between preference and discrimination and which can be dangerous if employed by AI in match-making processes. This chapter will also explore if there is a line that we shouldn't cross when it comes to how AI is used in our love lives? What are the ethical and moral implications for AI involved in human matchmaking?

So far, AI dating services are all *voluntary,* meaning the user can opt out of using online dating platforms at any time, or not use them at all. But, even if they are voluntary and optional,

[1] Cambridge Dictionary, "Meaning of "there are plenty of fish in the sea"", *Cambridge Dictionary,* 2019, https://dictionary.cambridge.org/us/dictionary/english/there-are-plenty-more-fish-in-the-sea

there should be consideration by the website and app developers about the ethical implications of their technology on individual users and humanity more broadly. If the AI is using filters for a user that actively perpetuate race and socio-economic bias in dating sites, does this help or hinder our society? The notion that our compatibility with someone *could* be calculated might also be enough to change how humans go about seeking love and partnerships in the future. Will the usage of AI interfere with human nature, and our own feelings and emotions? There's no question that technology and thus AI will continue to be used and expanded in the realm of love and dating. Technological advancements in online dating *have* the ability to provide a lot of ease to users, make finding a match straightforward without the hassle of scanning through multiple users and profiles. This chapter is not to dissuade people from using these apps, but rather to start a conversation on how AI can be used in a manner that benefits society instead of perpetuating certain biases. Focusing on the more popular and mainstream dating apps and websites that cater to finding love online, I will investigate the current technologies at play, as well as examining human bias and how humans go about selecting a potential mate, delving into the murky waters of AI involvement in Love.

Are Dating Apps Becoming "Essential" in Finding Modern Day Love?

For some, a question that may arise when the topic of the ethical nature of AI in dating is being discussed is, does it truly matter? Are people really using online dating at a rate where we need a discussion of the ethical implications of AI? An extensive guide of statistics was compiled in 2018 that looked at a plethora of data concerning online dating.[2] One of the statistics the article cites is that according to the Statistic Brain Research Institute, more than 49.7 million Americans have tried online dating, which isn't far from the total amount of single people in the U.S., at 54.4 million.[3] In addition, the data they compiled included a poll conducted by the Pew Research Center which showed that as of 2015, 59% of U.S. adults had an overall "positive attitude" about online dating.[4] Lastly, one of the most profound statistics the article shared was that one in five committed relationships started online.[5]

At least in America, online dating is not only becoming more and more used, but is being seen as a positive option in finding a committed partner. If these trends are to continue in an upward projection, there likely will be a time when online dating might be one of the most popular ways in meeting a future partner. With this in mind, the implications of AI being implemented in these dating sites and apps does in fact need to be discussed. If AI continues to be used without public knowledge and discourse in how it is implemented, bias can and will continue to be present towards certain users and demographics. Grindr, a popular dating app for people (mostly men) who are looking for same-sex relationships has already been accused of perpetuating racial discrimination in its usership, in part due to their being a racial preference filter users can employ.[6] One anecdote that was posted on the National LGBTQ Task Force

---

[2] Hayley Matthews, "27 Online Dating Statistics & What They Mean for the Future of Dating", *Dating News,* June 15th 2018, https://www.datingnews.com/industry-trends/online-dating-statistics-what-they-mean-for-future/.
[3] Ibid.
[4] Ibid.
[5] Ibid.
[6] Chris Stokel-Walker, "Why is it OK for online daters to block whole ethnic groups?", *The Guardian,* September 29th, 2018.

website cited a user who is Black but would identify himself as "mixed" on Grindr because he would get more "swipes" on his profile. The same user recalled grossly offensive comments he received when he did have his identity set to "Black", with one white-identifying man asking if he, "wanted to make a white man his slave?"[7]

It is ethically wrong to continue to supply AI that perpetuates racial and other discrimination without at least user's full knowledge and consent. Users of minority races and identities should be aware that by signing up for online dating, the app might in fact negatively value them and give them far fewer profiles to review than other non-minority users have to choose from. If dating apps become one of the sole methods by which singles meet other singles, to negatively discriminate against certain populations will have an overall detrimental effect to that minority or group's birth rates, success in love, and overall happiness. Discussing the ethical and moral implications of AI now in dating sites and apps before it becomes more widespread will at minimum start the conversation on how best to provide the best user experience for all demographics of people.

While the use of AI in dating apps raises questions about racial discrimination, overall dating apps themselves actually better the chances of interracial relationships. There is evidence that online dating is changing the demographics of serious relationships and marriages that are being formed, with a rise in interracial relationships specifically as a new trend. Online dating sites and apps are not the issue, but rather the AI and the algorithms being employed within the dating apps to scan and select profiles for users. An article from the MIT Technology Review cited new research by researchers from the University of Essex and University of Vienna that showed a steady increase in interracial marriages in spikes and leaps concurrent with the invention of online dating in 1995 with Match.com, and the most recent spike in interracial marriage in 2014 with the creation and use of Tinder for smartphones.[8] The research came to these conclusions through stimulating what happens when extra social links are added to an individual's network. Individuals are far more likely to date someone they know through loose social ties like a friend or professional circle than a complete stranger.[9] This is because these are the people they physically know, have interacted with, and have been able to have in-depth conversations with. Through online dating platforms an individual technically has the chance to meet someone and develop these loose social ties with a person they may have never met otherwise, who wouldn't have been in their original friend or social circles. For individuals who have fairly homogenous friend and professional circles, to date someone outside their race would be unlikely because where would they get the chance to meet them?

The probability of meeting and matching with someone from a different race and ethnicity increases when people use online platforms that allow them to interact with people

---

https://www.theguardian.com/technology/2018/sep/29/wltm-colour-blind-dating-app-racial-discrimination-grindr-tinder-algorithm-racism

[7] Rick Mula, "Wonky Wednesday: Racism in Gay Online Dating", *National LGBTQ Task Force,* 2019, http://www.thetaskforce.org/wonky-wednesday-racism-in-gay-online-dating.

[8] Emerging Technology from the arXIV, *First Evidence that Online Dating Is Changing the Nature of Society,* MIT Technology Review, October 10, 2017, https://www.technologyreview.com/s/609091/first-evidence-that-online-dating-is-changing-the-nature-of-society/.

[9] Emerging Technology from the arXIV, *First Evidence that Online Dating Is Changing the Nature of Society,* MIT Technology Review, October 10, 2017, https://www.technologyreview.com/s/609091/first-evidence-that-online-dating-is-changing-the-nature-of-society/.

outside of their current social circle.[10] The MIT Technology review research also predicted that marriages that were created through online dating tend to be stronger than those from traditional methods of courtship, having lower rates of marital breakup overall.[11] Would the implementation of AI in these dating apps further increase the growth of interracial marriages or would it decrease them? Depending on what the AI uses to filter matches and how it was programmed to view individuals' beauty, its implementation in dating apps might decrease the growth of interracial marriage. If the AI is taught by its programmers that certain features, or races are more appealing/more "beautiful" than others it will have a negative effect on people who don't fit those standard. What will be talked about later on in the chapter is the implementation of bias involved with AI. For example, using an algorithm that defines beauty with European standards such as light or fair skin, Eurocentric features and hair will ultimately filter out individuals who are not from this background. This leaves racial minorities at least in the U.S. at a clear disadvantage for meeting potential matches online. If the AI is using algorithms to filter potential matches for us to view, is there a need to ensure that racial bias isn't programmed into the algorithm? This investigation of bias in AI will be explored further.

What is Online Dating Anyway and the Current Landscape

Whether we want to admit it or not, online dating is here to stay. With platforms from websites like Match.com, and eHarmony, to more modern phone applications like Tinder, Bumble, and Hinge, electronic ways of meeting available singles aren't groundbreaking anymore. In fact, according to a popular wedding planning site, The Knot, a 2017 survey they conducted found that online dating was the most popular way that currently engaged couples had met.[12] In the beginning stages of online dating, users on Match would submit an electronic ad for themselves for other singles to search through before selecting one to which to send a message.[13] Now, almost all dating sites and apps are implementing algorithms to help narrow your choices.[14] These algorithms, which are unique to each different dating site or app, gather data on how users interact in order to calculate which profiles will then appear in your feed to "swipe on" or as potential matches on the more traditional sites to view.[15] Yet, the future will take these algorithms and AI to a completely new level of interaction and involvement in our dating and love lives. With the possibility of these sites employing AI to further the selectivity of the profiles and fellow users we get to view, they are coming out with a whole other onslaught of features including suggesting where you should take your potential match on a first date, conversation topics, and informing users on the projected duration of a relationship.[16]

---

[10] Ibid.
[11] Ibid.
[12] The Knot, "The Knot 2017 Jewelry and Engagement Study", Nov. 9th 2017, https://www.prnewswire.com/news-releases/only-1-in-3-us-marriage-proposals-are-a-surprise-engagement-ring-spend-rises-according-to-the-knot-2017-jewelry--engagement-study-300552669.html, March 16th 2019.
[13] Ally Marotti, "Algorithms behind Tinder, Hinge, and other dating apps control your love life. Here's how to navigate them", *Chicago Tribune,* December 6th 2018.
[14] Ibid.
[15] Ibid.
[16] Basile Dekonink, "For Dating Sites, Artificial Intelligence V. The Human Heart", *WorldCrunch,* July 19th 2018, https://www.worldcrunch.com/tech-science/for-dating-sites-artificial-intelligence-v-the-human-heart.

While no individual has ever been *required* to subscribe or use a dating app, these services' rising popularity means that more and more people might *feel the need* to engage in online dating either in addition to or replacing traditional methods. If the current single population is using online platforms as their main method of finding dates, the traditional way of going to bars or striking up a conversation in public might not only be unsuccessful, but also out of the norm. If the majority of singles in your area are not going out in "real" spaces to find a potential date, then going to a bar with friends might only result in seeing a lot of couples on their first date from meeting online, and less people also searching for someone to mingle with. Using dating sites and apps may in fact be the best method to obtain a date and relationship in the nearby future. With this rise in popularity of online dating, there needs to be a more careful observation of what the technology of these apps does and does not do, and how AI plays a factor in your dating process. Later on in the chapter I delve into considering implementing legal measures to ensure that these apps are not unfairly discriminating groups of people, this would have the government ultimately be the enforcer to ensure that the technology is fair. The companies and programmers who are creating the AI for these sites should think critically about how their technology is being implemented.

Tech companies should have a moral and ethical responsibility to ensure that their apps and websites and the AI currently in use within them isn't causing systematic harm to certain demographics of people. Because of how personal an individual's love life and relationship status can be, a negative experience on a dating platform can severely impact the confidence, and security of a user. This is particularly harmful if the app negatively values certain identity facets like skin color and race, or class status. The programmers who create the AI and program the algorithms for match selections need to consider and evaluate just *how* the technology filters the users as well as make sure the users are aware of exactly how the AI functions in the dating apps. If they don't, not only will some users have a negative experience on these sites, but also users *won't be aware* of how these apps function, and the filters they employ. An evasion of the problem could be had if the apps maintain at a level of use that is relatively optional for single adults to use, but not *necessary* for them to find love. Then, for those who use the apps, ethical challenges surrounding the usage of AI could be solved with a simple user consent button that explains that the apps implement algorithms which filter users on a variety of things and data within their profile. These apps currently don't tell you the filters and algorithms being employed to find you "your perfect match", yet if they did, would their consent be enough?

Individuals who are using these apps know that they are filtering the distance away, and the age of people they wish to view when setting up their profile on apps like Tinder. What is less transparent is why certain profiles appear first or more frequently on their feed than others, or that this even happens. An individual doesn't necessarily need to know the exact technology at play in their dating apps, but there should be user education available that lets them know that all of these sites run off of *algorithms* that select specific profiles for them to view and hide other profiles from them as a result. Tinder uses an "Elo Score" in the app, a system used first by competitive chess players to rank their skill level. [17] When used in Tinder, the Elo Score rates a user's "desirability" which is compiled of how many "swipes" a user profile gets from

[17] Austin Carr, "I Found Out my Secret Tinder Rating and Now I wish I Hadn't", *Fast Company,* January 1st 2016, https://www.fastcompany.com/3054871/whats-your-tinder-score-inside-the-apps-internal-ranking-system.

individuals (how often someone wants to match with that user), as well as the "desirability" of the users they "match with" (when both profiles swipe on each other's profile).[18]

Tinder is one of the simpler dating sites, asking for users to set a limit for the age range of people they would like to match with and distance they are willing to travel to meet someone. There are other, more complicated apps like Bumble, which allows someone to input their religion, political preferences, and level of seriousness in dating that they are looking for.[19] There are still more apps and sites that tailor to varying age ranges, regional and global demographics, sexual preferences and more. For the apps using more and more preferences in their algorithms – some even going as far as asking for racial preferences – one could see how these filters could lead to discrimination for certain groups.[20] At the very least, if users could click a "confirmation of use" button before starting a subscription where the user must signify that they understand that algorithms and filters are being employed to give them a selection of the available users on the app, *and* they can opt out of these filters being employed, would this solve the issue? It might potentially solve the problem of the programming and app developers being held responsible for sustaining bias and prejudice in online dating, but it does little to curb or prevent what individual *users* are actively engaging in discrimination and bias with the help of the AI technology. Furthermore, if the use of these dating apps and sites to find a potential partner becomes unavoidable in some populations because these services are the only effective method to modern day dating, then the ethical implications of AI need to be explored further on an individual basis.

Bias: In Humans and in AI

With one of the prominent features of using AI in dating sites being that it would filter and select a predefined set of profiles for a user to view and "match" with, the issue of bias comes at play. All humans have bias, whether we want to admit it or not. In online dating, a recent study of over a million users of an online dating site showed that, "Black individuals were 10 times more likely to contact Whites than Whites were to contact Blacks."[21] This study was performed with over a million nationwide users of an online dating site that analyzed their personal online profiles, but also records of communication between matches in the U.S. In addition, in 2014, OkCupid found through compiling data of successful matches that Black women and Asian men were more likely to be rated far lower than other ethnic groups on their dating site.[22] While some can argue that a person's *preference* is different than being *prejudiced* towards certain groups, this is very "grey" territory. There is inherently little harm in genuinely being attracted to certain features or characteristics of people; preferences help us narrow our

---

[18] Ibid.

[19] Sarah Perez, *Bumble now lets you filter potential matches on Bumble Date, Bizz and BFF,* Tech Crunch, January 2019, https://techcrunch.com/2018/12/18/bumble-now-lets-you-filter-potential-matches-on-bumble-date-bizz-and-bff/.

[20] Chris Stokel-Walker, "Why is it OK for online daters to block whole ethnic groups?", *The Guardian,* September 29th, 2018, https://www.theguardian.com/technology/2018/sep/29/wltm-colour-blind-dating-app-racial-discrimination-grindr-tinder-algorithm-racism

[21] Gerald Mendelsohn et al., Black/White dating online: Interracial courtship in the 21st century, Psychology of Popular Media Culture, Vol3(1), American Psychological Association, January 2014.

[22] OkCupid, *Race and Attraction 2009-2014,* OkCupid Blog, Sept 10th, 2014, https://theblog.okcupid.com/race-and-attraction-2009-2014-107dcbb4f060.

choice of a potential partner, and if we didn't have preferences of any kind there would in fact be *far* too many "fish in the sea" for us to date. What becomes problematic is when users fall into racial and stereotypical tropes with their preferences, leading to fetishizing of certain groups as well as solely seeking only those groups to romantically pursue and on the opposite spectrum unfairly rejecting whole groups of people. Preferences do not have to be necessarily harmful on an individual basis, but when it is on a macro level where they are formed historically and culturally due to racial oppression, slavery, and segregation it becomes prejudice and discriminatory on a societal level which is where harm comes in.

In using AI to filter matches does this give way to users' preferences, or their prejudice? Can one say with full confidence that someone's racial preference in dating solely within their race, one particular race, or choosing to not even entertain the possibility of dating a particular racial group is purely neutral? In the context of the U.S. history of racism, colorism, and discrimination against certain racial groups I believe it is better for race relations within the U.S. that dating sites and apps don't employ racial filters on their apps. This is *because* it is so difficult to determine if someone wants to date a certain racial group based on preferences versus based on prejudice. What programmers and creators of AI technology within dating sites should strive for is creating AI that doesn't rely on racial or class filters in determining users "perfect match". I argue that one could be more open to meeting other individuals from different backgrounds (racially, socially, politically, and religiously) on online dating if the AI uses filters like matching based on personality trait or other "neutral" filters. If dating sites were to do this, AI could work to stop some of the inherent racial and SES bias that persists in dating sites and apps by bolstering users' confidence that people of different demographics can be better matches for them then they previously believed.

The possibility that AI can combat our internal prejudice with these dating apps make it seem like an extremely positive addition. However, there is a need to discuss how AI can also be biased when filtering users. The data we saw earlier showed that due to online dating, interracial relationships have been on the rise. Yet, with AI being more and more implemented into these dating sites, there needs to be consideration into just *how* the system is filtering individuals and what sort of filters it is implementing. If for example, the AI were to create its algorithms for these dating sites using the method of a *deep-learning model,* then it would be fed a high volume of data for it to start to recognize patterns and learn.[23] Deep-learning models are when AI are programmed by given large amounts of real-world data to "digest" and then find patterns and similarities so that it can create an algorithm to implement in systems. If the AI in dating sites and apps were fed magazine images of women to filter for attractiveness, it would end up with a high ratio of fair/light skinned women as opposed to women with darker skin tones. In 2017, an article researched that out of 442 times a model, actress, television personality, singer, or even a political figure graced the cover of Vogue, a popular fashion magazine over a 30 year span, from 1980's to 2017, only 30 of those times featured a black or African-American person.[24] This is less than 10% of the entire Vogue magazine publications. This problem could be balanced by having the AI also digest minority magazines, however that would have to be an *intentional*

---

[23] Karen Hao, *This is how AI bias really happens-and why it's so hard to fix,* MIT Technology Review, February 4th 2019, https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/.
[24] Tori Owens, *As America grows more diverse, fashion magazine covers are slow to show progress,* Peninsula Press, Stanford University, August 25th 2017, http://peninsulapress.com/2017/08/25/fashion-magazine-covers-diversity-analysis/.

move done by the programmers as the "default" "mainstream" media is usually coded as Eurocentric. Similar fashion and beauty magazines also follow this dangerous trend making the use of deep-learning for AI in these dating sites potentially harmful for users who have a darker complexion if the AI is gathering data from mainstream news sources and publications.

What are the solutions then to bias in deep-learning, if the very data we use to teach the AI is itself skewed? One of the most obvious examples is to ensure the data and information being used during deep-learning offers a wide and diverse range of individuals when searching for *"attractiveness",* varying *"personality types*", or other filters the dating site may employ. In fact, certain dating apps might find more success and higher user-ship if they were to only offer filters based on personality traits and values and not physical appearance. This would require the creation of the AI and the programming by the companies to be more thought out and intentional. Companies would need to specify that the AI use deep-learning that pulls from global media and news sources and is not based solely on U.S. or euro-centric data. While we couldn't change or stop how the users themselves "match" with potential partners when using the app or site, we could prevent some unneeded marginalization by ensuring that the algorithms the AI is using aren't racially skewed or biased towards certain SES and ethnic groups. One answer that is apparent is that we can't *not* filter an individual's potential matches.

With the rise and increase in users, online and app dating will only continue to grow. With a potential for a dating pool to be in the thousands, there will be a *need* for some level of filtration in order for individuals to have the potential to match with someone without being overwhelmed by their choices and ultimately getting off the site before matching with anyone. While some dating sites cater towards specific groups and demographics, the more general dating sites like Match, EHarmony, and apps like Tinder and Bumble to name only a few should be transparent in the ways in which they filter individuals for users to view. Transparency of these processes would give users more knowledge into what exactly is being calculated and observed when they create an online dating profile, and likelihood of them to receive high versus low number of matches.

Another question that needs to be considered is, what if people want these racial and socioeconomic preferences as options for filtering? People can date whom they want to date, and love whom they want to love, at least in the U.S. Should users be allowed to act on their biases or should AI be created that intentionally doesn't allow for filters that employ any type of racial, socioeconomic or demographic marker on the person? Users ultimately have the choice on who they "swipe" or want to connect with, the app and site merely giving them the selection of potential matches to view. A solution that might ease the minds of some is for apps to be transparent in the filters they use to select the individual profiles it shows the viewer. This may lead to apps and sites who intentionally don't want any filters to be associated with their site, and apps and sites that employ every possible form of filter. Morally, it would be best if a user could pick a site with enough knowledge of what that site does and how it operates so that they are well-informed of their chances for a match on certain sites or apps they are interested in joining. Societally, apps that employ filters of race, class, or socioeconomic status shouldn't be accepted as "morally legitimate" because that is perpetuating blatant discrimination and fetishizing of certain groups of people. We can't at this stage ban apps from employing these filters: it isn't illegal for them to do so. But we can try to inform users that by employing these filters on dating sites and websites we are negatively impacting the overall dating community and experience for certain groups of people and perpetuating hate and bias. Ultimately the user will have final say

on *whom* they end up matching, but AI can do its part to ensure the range of potential matches a user sees is both socio-economically and racially diverse.

AI's Place in the Law?

In thinking about regulations, an investigation into the legality of AI and love need next be examined. Ethics and the law are different, as something can be ethically wrong like not tipping a waiter at a restaurant in the U.S. with the previous knowledge that they survive off of tips due to how the wage system of waitressing works but it isn't illegal. On the other end of the spectrum, causing undue harm or pain to someone, or even killing someone is both ethically and morally wrong, as well as legally. Currently there are very limited moves to legally regulating the usage of AI in the U.S. The Obama administration started to create a task force centralized on AI regulations and research, putting out two reports centered around establishing an ethical AI policy, and then the effects of AI-driven automation.[25] However, the task force was shut down during the Trump Administration and the reports largely dismissed.[26]

In looking at laws actually in place regarding efforts against discrimination, those too need more attention. Current legislation that focuses on discrimination makes no mention of tech companies, programmers, or non-human (AI) as subjects capable of discrimination, nor an online platform as a place where discrimination could occur. Where online dating regulation could be improved upon and implemented is within the expansion of the U.S. Code 42 concerning public accommodations.[27] One clause titled "Equal Access" states that "All persons shall be entitled to the full and equal enjoyment of the goods, services, facilities, privileges, advantages and accommodations of any place of public accommodation".[28] These places of "public accommodation" include restaurants, movie theaters, concert halls, so why can't they include online?

Online dating sites and apps are replacing the traditional method of going out to meet potential dates and people, they are just as much a public space as somewhere in person, the only difference is that they are virtual. People are either signing up to use a free app, or they are spending their money in order to enter the space of the online site or application. If dating in this age is going to start largely on an online platform, shouldn't these online spaces be categorized as a place of public accommodation? Morally it is wrong to cause harm to someone, and discrimination is one of the means in which harm can be done to an individual. If we discriminate on people on the basis of things they can't control like their race, ethnicity, or gender when they are at a public place, it is both legally and morally wrong. I argue that this should be implied for discrimination in online dating sites and love. By legally requiring sites and apps to have AI filters and algorithms that have been intentionally created to not be biased or discriminatory towards certain demographics of people, it would ethically be better in preventing undue harm, but also could be enforced legally.

---

[25] Melody Guan, *Regulating AI in the Era of Big Tech,* The Gradient, July 8th 2018, https://thegradient.pub/regulating-ai-in-the-era-of-big-tech/.
[26] Ibid.
[27] Cornell Law School, *Places of Public Accommodation,* U.S. Code 42 THE PUBLIC HEALTH and WELFARE, Legal Information Institute, 1992, https://www.law.cornell.edu/uscode/text/42/2000a.
[28] Ibid.

If we are to include online platforms like dating sites as a place of public accommodation and thus protected against undue discrimination, this might open up a massive can of worms within society. What other sites could be protected? Would a user have to pay for a service (paying for a dating membership or paying for a premium user-ship in one of the dating apps) in order for their time on the app to qualify as being at a place of public accommodation, much like ordering at a restaurant or paying for a movie theater ticket? If they can use a free service like Tinder, wouldn't this be the same as attending a free public event at a park or community center? What about other non-dating sites and the usage of AI technology to filter results like in Facebook? The scope of those questions seem too big for the purposes of this chapter, but they are something to consider when evaluating if legal protection needs to be in place with online sites and platforms.

The second question that also needs to be considered, is who will ultimately be held responsible for the discrimination from the AI algorithms? Would it be the company who employs this technology, or the engineers and computer scientists who create the deep-learning models? Depending on who would be held accountable, further questions and regulations would need to be put in place to ensure that the AI being implemented is creating the best quality system, while also not falling into the realm of undue discrimination. The legality of AI is just beginning, and I argue that legality and morality are heavily intertwined, oftentimes our legal code is merely based off of a certain standard of ethics or morals in our society that has been employed for generations, and now is being enforced through legal means. While no laws have been implemented yet, the likelihood of legislation expansion concerning the use of AI is high. Society as a whole and specific legal systems must first identify if we consider the Internet to be a place of public accommodation, then questions and laws concerning dating applications and sites may be considered. In this regard, a further examination of anti-discrimination laws and the intersection of the law and morals concerning AI will be needed further down the line as the technology becomes more and more used.

Conclusion

So is AI in Love unethical? It depends on how it is used. The rise in the usage of online dating and apps makes the likelihood of this problem going away slim. The implications of AI being implemented with the potential for bias within online dating means there needs to be a closer look into how the algorithms are created and the way deep-learning is implemented. I argue that we need to hold programmers and tech companies accountable for the products they create and the AI they employ. It is not equal responsibility for the programmers and the companies, the companies who are ultimately producing these apps and especially the ones gaining money off of these sites are more accountable than the programmers who are simply hired by these larger companies to implement the AI. It could be impactful to have programmers participate in anti-bias trainings before they create the AI and the algorithms being used. As a whole, it is ethically wrong to discriminate against someone on the basis of race, sex, or nationality, and if dating sites and apps are doing this it should be both an ethical and legal question to consider. Dating sites and apps have the potential to do a lot of good for society, increasing the population of long-term couples and helping narrow down potential choices of matches to a tolerable amount of profiles an individual can view without feeling overwhelmed. There should be transparency on app and websites about the AI being implemented, and how

matches are being filtered. Users should be able to go on these dating sites and feel like they have a fair shot at getting and receiving matches, regardless of their demographics. There can indeed be, “too many fish in the sea” yet with AI being used properly and fairly in dating sites and apps we can feel confident knowing we are seeing an equal representation of the available dates in our area without being drowned in options and choices.

Works Cited

"As America Grows More Diverse, Fashion Magazine Covers Are Slow to Show Progress." Peninsula Press. October 25, 2017. Accessed March 21, 2019. http://peninsulapress.com/2017/08/25/fashion-magazine-covers-diversity-analysis/.

Carr, Austin, and Austin Carr. "I Found Out My Secret Internal Tinder Rating And Now I Wish I Hadn't." Fast Company. May 10, 2017. Accessed April 07, 2019. https://www.fastcompany.com/3054871/whats-your-tinder-score-inside-the-apps-internal-ranking-system.

Cornell Law. "42 U.S. Code § 2000a - Prohibition against Discrimination or Segregation in Places of Public Accommodation." Legal Information Institute. Accessed March 20, 2019. https://www.law.cornell.edu/uscode/text/42/2000a.

"Dating Apps Use Artificial Intelligence to Help Search for Love." Phys.org - News and Articles on Science and Technology. November 8, 2018. Accessed March 21, 2019. https://phys.org/news/2018-11-dating-apps-artificial-intelligence.html.

Greenacre, Martin. "For Dating Sites, Artificial Intelligence v. The Human Heart." Worldcrunch. July 19, 2018. Accessed March 17, 2019. https://www.worldcrunch.com/tech-science/for-dating-sites-artificial-intelligence-v-the-human-heart.

Guan, Melody. "Regulating AI in the Era of Big Tech." The Gradient. July 09, 2018. Accessed March 21, 2019. https://thegradient.pub/regulating-ai-in-the-era-of-big-tech/.

Hao, Karen, and Karen Hao. "This Is How AI Bias Really Happens-and Why It's so Hard to Fix." MIT Technology Review. February 04, 2019. Accessed March 17, 2019. https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix/.

Knot, The. "Only 1 in 3 US Marriage Proposals Are a Surprise; Engagement Ring Spend Rises, According to The Knot 2017 Jewelry & Engagement Study." PR Newswire: Press Release Distribution, Targeting, Monitoring and Marketing. November 09, 2017. Accessed March 17, 2019. https://www.prnewswire.com/news-releases/only-1-in-3-us-marriage-proposals-are-a-surprise-engagement-ring-spend-rises-according-to-the-knot-2017-jewelry--engagement-study-300552669.html.

McMullan, Thomas. "Are the Algorithms That Power Dating Apps Racially Biased?" WIRED. February 17, 2019. Accessed March 21, 2019. https://www.wired.co.uk/article/racial-bias-dating-apps.

Mendelsohn, Gerald A., Shaw Taylor, Lindsay Fiore, and Andrew T. Cheshire. "Black/White Dating Online: Interracial Courtship in the 21st Century." American Psychological Association. 2018. Accessed March 17, 2019. https://psycnet.apa.org/record/2014-02726-002.

Mula, Rick. “Wonky Wednesday: Racism in Gay Online Dating.” National LGBTQ Task Force.2019.Accessed April 20, 2019. http://www.thetaskforce.org/wonky-wednesday-racism-in-gay-online-dating/

OkCupid. "Race and Attraction, 2009–2014." The OkCupid Blog. September 10, 2014. Accessed March 19, 2019. https://theblog.okcupid.com/race-and-attraction-2009-2014-107dcbb4f060.

Perez, Sarah, and Sarah Perez. "Bumble Now Lets You Filter Potential Matches on Bumble Date, Bizz and BFF." Bumble Now Lets You Filter Potential Matches on Bumble Date Bizz and Bff/. December 18, 2018. Accessed March 20, 2019. https://techcrunch.com/2018/12/18/bumble-now-lets-you-filter-potential-matches-on-bumble-date-bizz-and-bff/.

Stokel-Walker, Chris. "Why Is It OK for Online Daters to Block Whole Ethnic Groups?" The Guardian. September 29, 2018. Accessed April 07, 2019. https://www.theguardian.com/technology/2018/sep/29/wltm-colour-blind-dating-app-racial-discrimination-grindr-tinder-algorithm-racism.

"THERE ARE PLENTY MORE FISH IN THE SEA | Definition in the Cambridge English Dictionary." THERE ARE PLENTY MORE FISH IN THE SEA | Definition in the Cambridge English Dictionary. Accessed April 07, 2019. https://dictionary.cambridge.org/us/dictionary/english/there-are-plenty-more-fish-in-the-sea.

US Census Bureau. "America's Families and Living Arrangements: 2016." America's Families and Living Arrangements: 2016. May 04, 2018. Accessed April 07, 2019. https://www.census.gov/data/tables/2016/demo/families/cps-2016.html.

Assessing Facial Recognition: An Ethical Exploration

John Benhart

Introduction

"Without a thoughtful approach, public authorities may rely on flawed or biased technological approaches to decide who to track, investigate or even arrest for a crime. Governments may monitor the exercise of political and other public activities in ways that conflict with longstanding expectations in democratic societies, chilling citizens' willingness to turn out for political events and undermining our core freedoms of assembly and expression. Similarly, companies may use facial recognition to make decisions without human intervention that affect our eligibility for credit, jobs or purchases. All these scenarios raise important questions of privacy, free speech, freedom of association and even life and liberty." – Brad Smith, President of Microsoft[29]

***

In July 2018, Microsoft President Brad Smith released a report calling for congressional regulation of facial recognition.[30] Microsoft stands to gain much from the technology, which consists of software and algorithms that can be used to identify faces in images and video. Practical applications of facial recognition include cell phone cameras, hundreds of photos in a Facebook photo album, and crime surveillance in a local park. The accelerating development of facial recognition could enhance Microsoft's products and add new capabilities that generate profit for the company. Moreover, facial recognition has numerous beneficial applications that could help society at large, such as more personalized advertising and retail experiences, improved security and law enforcement, and streamlined, personalized use of certain services like Facebook's photo-tagging service. However, in his report, Smith highlights that, despite these economic and societal benefits, massive risks underly facial recognition's continued unregulated (and at times unethical) development. He lays out his concerns above, citing issues of "privacy, free speech, freedom of association and even life and liberty" that arise in situations of improper use by governments and companies alike. Moreover, Smith acknowledges that large tech firms producing the tech will not be enough to control facial recognition; it will need to be elected officials who do the job.

Why would one of the biggest players, who stands to gain substantially from growth in facial recognition, push for regulation of facial recognition? Of course, Microsoft can gain positive publicity by taking a stand for issues such as privacy, as other large tech firms have done. But, in December 2018, another statement followed up the initial message from Smith, calling for regulation and laying out a set of guiding ethical principles.[31] These additional steps

---

[29] Brad Smith, "Facial recognition technology: The need for public regulation and corporate responsibility," *Microsoft*, last modified July 13, 2018, https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/.

[30] Ibid.

[31] Brad Smith, "Facial recognition: It's time for action," *Microsoft*, last modified December 6, 2018 https://blogs.microsoft.com/on-the-issues/2018/12/06/facial-recognition-its-time-for-action/.

go beyond publicity to advocate for specific legislation. Microsoft sees regulation as inevitable and would prefer to shape the laws that will restrict the technology, guaranteeing that its competitors will also be restrained by the same ethical principles which Microsoft upholds. In the meantime, Microsoft has denied contracts that it deems to raise human rights concerns. In early 2019, Microsoft rejected contracts with California law enforcement, which wanted to scan the face of anyone pulled over at a traffic stop, and with an anonymous country "that the nonprofit Freedom House had deemed not free."[32] The country had wanted to install the technology widely in its capital city, which could be used to violate citizens' liberty. As Microsoft's statements and actions demonstrate, facial recognition presents urgent ethical dilemmas that need to be considered now.

Facial recognition technology has been around since the 1980s, though the technology has become dramatically more sophisticated in recent years. Artificial Intelligence (AI) and Machine Learning (ML) have coupled with Big Data techniques to make marked performance improvements in facial recognition. An article produced by Gemalto—a company marketing its own security and facial recognition solutions—notes that the performance of facial recognition technology has matured enough that in 2014, Facebook achieved a 97.25% accuracy rate for its DeepFace algorithm, which determines whether two faces belong to the same person.[33] The human accuracy rate for this test was 97.53%; so, DeepFace performed almost as well as humans. Just a year later, Google achieved an accuracy rate of 100% when using its FaceNet algorithm on the "Labeled Faces in the Wild" database of over 13,000 headshot images of celebrities and other well-known people.[34] By now, these results are more than three years old, and the underlying artificial intelligence supporting the algorithms continues to rapidly develop.

Such accurate facial recognition algorithms can be used in many ways. The technology is now being marketed and deployed in various business applications, including retail, marketing, health, law enforcement, and security to improve safety.[35] However, existing manifestations and possible future iterations of facial recognition technology raise important issues regarding discrimination, privacy, government surveillance, and others. Enough concerns persist that two of the biggest players in facial recognition technology, Microsoft and Amazon, are pressing for government legislation "that protects individual civil rights and ensures that governments are transparent in their use of facial recognition technology."[36] This desire for regulation of their own high-growth businesses indicates willingness to ensure that the industry and its clients do not go too far with the technology before Congress decides on its proper use. Moreover, it recognizes that for them, questions of ethical implementations are not a thought experiment but are in fact

---

[32] Joseph Menn, "Microsoft turned down facial-recognition sales on human rights concerns," *Reuters*, last modified April 16, 2019, https://www.reuters.com/article/us-microsoft-ai/microsoft-turned-down-facial-recognition-sales-on-human-rights-concerns-idUSKCN1RS2FV.

[33] "The top 7 trends for facial recognition in 2019," *Gemalto*, accessed April 20, 2019, https://www.gemalto.com/govt/biometrics/facial-recognition.

[34] "Labeled Faces in the Wild Home," *University of Massachusetts – Amherst*, last modified January 1, 2018, http://vis-www.cs.umass.edu/lfw/.

[35] "The top 7 trends for facial recognition," *Gemalto*.

[36] Tom Simonite, "Amazon Joins Microsoft's Call for Rules on Facial Recognition," *Wired*, last modified February 7, 2019, https://www.wired.com/story/amazon-joins-microsofts-call-rules-facial-recognition/.

immediate considerations as they and their competitors alike assist clients in applying facial recognition in the real world.[37]

Microsoft's July 2018 report recognizes the rapid shifts occurring in the industry and the need for ethics to be applied to answer some of the biggest questions. Questions Microsoft proposes span possible technical applications and ethical concerns:[38]

- "Should use of facial recognition by public authorities or others be subject to minimum performance levels on accuracy?"
- "Should law enforcement be subject to human oversight and controls, including restrictions on the use of unaided facial recognition technology as evidence of an individual's guilt or innocence of a crime?"
- "Should the law require that companies obtain prior consent before collecting individuals' images for facial recognition?"
- "Should we create processes that afford legal rights to individuals who believe they have been misidentified by a facial recognition system?"

The above questions address numerous possible ethical difficulties that could arise from facial recognition technology, including applications in law enforcement to apprehend potential criminals and potential nonconsensual use of data to train the algorithms. In this chapter, I look to answer a few of these questions and ask many more. First, I will elaborate on the current widespread uses for facial recognition technology and demonstrate the potential for positive implementation. Then, I will consider some key ethical questions that have arisen in already-controversial deployments. Finally, I will propose some possible ethical frameworks, consider the principles supported by Microsoft, and detail regulation currently progressing through the legislative process.

Facial Recognition: Artificial Intelligence, Big Data, and the Seeds of Controversy

Facial recognition technology is in flux. New techniques developed in the past few years have contributed substantial performance gains. These have resulted from the changing nature of the algorithms that power facial recognition, including a pivotal reliance on artificial intelligence (AI) as a key component. Most generally, facial recognition algorithms are those that identify and/or verify faces in images and video. Identification is the process of recognizing a face in a pre-existing database of faces; while verification (authentication) is the process of confirming one's identity using pre-recorded biometric data of their face.[39] Below the surface, however, the traditional facial recognition process is comprised of several subroutines which together constitute the algorithm's general pipeline.[40] First, face detection occurs, whereby the faces are identified in the image. This step focuses the rest of the work of the algorithm, eliminating other non-face objects, such as cats, trees, or a blue sky. Next, the image goes through the pre-processing phase, where images are standardized by number and dimensions of pixels so that

---

[37] Joseph Menn, "Microsoft turned down facial-recognition".

[38] Brad Smith, "Facial recognition technology".

[39] Christopher S. Milligan, "Facial Recognition Technology, Video Surveillance, and Privacy," *Southern California Interdisciplinary Law Journal*, 1 no. 9, (1999): 306, accessed April 20, 2019, https://heinonline.org/HOL/Page?handle=hein.journals/scid9&div=7&g_sent=1&casa_token=&collection=journals.

[40] Yaroslav Kuflinski, "How Facial Recognition Works," *Iflexion*, last modified July 17, 2018, https://www.iflexion.com/blog/face-recognition-algorithms-work/.

subsequent comparisons are accurate. After the images are normalized, feature extraction isolates the given features relevant to a face, including the eyes and mouth, coloration, and relative distances. Finally, face recognition matches these features to another face existing in a reference database of faces.

Traditional algorithms such as Principal Component Analysis and Binary Pattern have employed the above techniques but have been recently surpassed by advancements in Machine Learning. These algorithms—deployed by leading companies and various researchers around the globe—use Neural Networks, an AI technique through which a network of nodes designed to observe different features of faces is trained to represent a face as a sequence of numerical data (a faceprint) unique to an individual.[41] From there, comparisons between faces are simple—the faceprint representations of two faces can be compared and identified as the same or not. Once trained, these algorithms execute in the order of seconds, and they vastly outperform traditional algorithms, achieving near-perfect scores when tested on new databases. To train a facial recognition algorithm based on Neural Networks, an initial database is used as a training set. In the training set, faces belonging to the same person are labeled. The more faces that are present in this database, the better the outcome. After training and optimization of the Neural Network, the algorithm is ready to be tested, and a test set of photos of new individuals' faces is compiled. Testing protocol can vary, but one method allows the algorithm to view a new database of faces, then identify subsequent input faces from the database. An accuracy rate can be determined by calculating the number of correct identifications by the algorithm.

Beyond the various possible algorithmic techniques and AI, facial recognition systems work symbiotically with large amounts of data, initially requiring data for development and later requiring it to facilitate collection and analysis. As is common with Machine Learning models, Neural Networks require large amounts of data for training. The best facial recognition algorithms are trained on millions of faces. So, to facilitate the use and development of these new algorithms, databases of faces must be assembled. Some databases of thousands of photos are curated by universities for testing and training purposes; however, the largest technology and social media companies can generate the most effective databases of millions faces using data from their millions of users. Likewise, these larger companies often achieve the highest accuracy rates for their facial recognition algorithms. An example of a popular university database used for training and testing is the "Labeled Faces in the Wild" database curated by the University of Massachusetts – Amherst.[42] It contains 13,233 images from 5,479 well-known people such as Britney Spears, Tiger Woods, and Tony Blair. Software companies that develop photo-sharing platforms already have a wealth of user-identified faces to use for this purpose. Additionally, companies like Ever AI have repurposed their existing photo apps as a database to train facial recognition algorithms.[43] Ever AI's app had collected an estimated 13 billion images, which could yield a vast database to use for training. Some companies encrypt these databases, though security concerns can persist. Once created and trained, facial recognition algorithms can be

[41] Mark Sutton, "Facial Recognition Technology: 10 buzzwords demystified," *ITP.net*, last modified November 23, 2018. http://www.itp.net/618391-facial-recognition-technology-10-buzzwords-demystified.

[42] "Labeled Faces in the Wild Home," *University of Massachusetts – Amherst*, last modified January 1, 2018, http://vis-www.cs.umass.edu/lfw/.

[43] Jeff John Roberts, "The Business of Your Face," *Fortune*, last modified March 27, 2019, http://fortune.com/longform/facial-recognition/.

utilized in any instance where images or video are generated, from a phone's camera to the public spaces in an airport.

The described structure of facial recognition systems inherently present ethical challenges. First, accuracy can be an issue. Some companies' algorithms contain biases that discriminate against minorities, leading to flawed, potentially destructive outcomes when implemented. Second, training requires databases of faces, and the efforts of companies to harvest faces for this purpose lead to potential privacy and consent violations. Finally, widespread deployment of facial recognition opens a potential Pandora's Box of issues, all enabled by the opportunities afforded by the artificial intelligence's speed and ease of use. When considering any tool with immense capabilities, relevant ethical considerations move beyond its development, to its eventual use. With algorithms that can accurately identify faces in just a few seconds, facial recognition serves as a valuable tool, but not without its own potential for corruption.

Applications of Facial Recognition

Facial recognition has many opportunities to create significant benefits for society. Continued development of these uses is paramount, given the technology's almost unparalleled utility. One of the most pervasive uses has been in law enforcement and monitoring. Beyond simple closed-circuit television (CCTV) technology for monitoring and older facial recognition technology, current algorithms using AI, such as FaceFirst's technology, have been deployed to identify past criminals and assist law enforcement in assessing risk in specific situations.[44] FaceFirst and facial recognition technology has been used by retailers to reduce theft by non-employees by 34% and to reduce violent incidents in stores by 91%.[45] Additionally, law enforcement and government can use facial recognition technology to identify missing or trafficked people among crowds in public spaces. In April of 2018, almost 3,000 missing children were found in New Delhi, India using facial recognition to identify the children using live video from the city's video cameras in public spaces.[46] These uses could potentially save tens of thousands around the world from smuggling or other related dangers.

Additionally, facial recognition can be used for security purposes to verify the identity of individuals. Facial recognition is currently used to verify users of phones and other device security in products sold by Apple, Microsoft, Samsung, and Google. Combined with other biometric data like fingerprints, this application has a bright future. Moreover, airports have begun to implement facial recognition for real-time passport verification at airports, leading to the apprehension of criminals who use fake passports. In August 2018, a man flying into the United States through Washington Dulles International Airport from Sao Paulo, Brazil was

---

[44] Jeff John Roberts, "The Business of Your Face".

[45] Jesse Davis West, "21 Amazing Uses for Face Recognition – Facial Recognition Use Cases," *FaceFirst*, last modified May 2, 2018, https://www.facefirst.com/blog/amazing-uses-for-face-recognition-facial-recognition-use-cases/.

[46] Anthony Culbertson, "Indian Police Trace 3,000 Missing Children in Just Four Days Using Facial Recognition Technology," *Independent*, last modified April 24, 2018, https://www.independent.co.uk/life-style/gadgets-and-tech/news/india-police-missing-children-facial-recognition-tech-trace-find-reunite-a8320406.html.

caught with a fake passport.[47] Real-time comparison of the man's face and his French passport revealed a discrepancy. Facial recognition had only been used at this airport for three days when it caught its first perpetrator, and it was believed that the man's passport might have passed human inspection if not for the facial recognition technology. Additionally, schools are being made safer using facial recognition technology, such as that deployed by RealNetworks.[48] The company's SAFR technology can identify people as they enter a school building, allowing for administrators to better monitor flow of visitors and students. Faces are registered with the system, then the system can allow certain adults, such as parents, or students, if the school is willing to register students' faces, to enter the school when they present their faces. The technology has a 99.8% accuracy rate.

Facial recognition has myriad uses beyond law enforcement and security. Facebook, Google Photos, and other applications use it to identify and automatically tag people in photos. In the health sphere, facial recognition has been used to diagnose difficult-to-identify diseases, like 22q11.2 Deletion Syndrome, or DiGeorge Syndrome.[49] DiGeorge Syndrome can modify facial features of patients, including the ears and mouth. The algorithms and software used to evaluate patients had meaningful diagnosing impact: "Sensitivity and specificity were greater than 96% for all populations." DiGeorge Syndrome can be difficult for doctors to identify on their own, given high comorbidity rates and the need for molecular testing. So, facial recognition could greatly streamline diagnosis. In the retail space, facial recognition technology like FaceFirst can be used to prevent theft. Walmart has used FaceFirst to identify potential shoplifters and other offenders registered in its database to confront them in real time. The technology can notify Walmart employees within "7 seconds."[50] The application does not record non-offenders faces; so, it can alleviate some concerns of privacy violations for non-offenders. Walmart has since discontinued its use of FaceFirst because of the cost, but the technology could be used in the future. Moreover, facial recognition can be used by retail and consumer goods companies for personalized advertising and shopping experiences. For example, facial recognition could identify important, frequent shoppers when they enter a store, allowing employees to take special care with the shopper. Individualized advertising within stores could also augment the shopping experience. Companies will continue to innovate and develop additional uses—the powers of profit will almost certainly lead to this. And, this will create many boons for society at large, through the various mentioned applications and others, leading to safer, more efficient law enforcement, retail, and medical systems.

Dangers, Ethics of Facial Recognition Technology

[47] Tom Costello and Ethan Sacks, "New facial recognition tech catches first impostor at D.C. airport," *NBC Universal*, last modified August 23, 2018, https://www.nbcnews.com/news/us-news/new-facial-recognition-tech-catches-first-impostor-d-c-airport-n903236.

[48] Issie Lapowsky, "Schools Can Now Get Facial Recognition Tech for Free. Should They?," *Wired*, last modified July 17, 2018, https://www.wired.com/story/realnetworks-facial-recognition-technology-schools/.

[49] Paul Kruszka et al., "22q11.2 deletion syndrome in diverse populations," *American Journal of Medical Genetics*, 173 no. 4, (2017): 879-888, accessed April 20, 2019, https://doi.org/10.1002/ajmg.a.38199.

[50] Jeff John Roberts, "Walmart's Use of Sci-fi Tech To Spot Shoplifters Raises Privacy Questions," *Fortune*, last modified November 9, 2015, http://fortune.com/2015/11/09/wal-mart-facial-recognition/.

While facial recognition carries serious benefits for society, as a highly potent technology it also poses serious ethical challenges. Many of these issues arise in the above implementations of facial recognition, which have already displayed significant benefits. First, consider the technology's implementation for security and law enforcement. These demand particular scrutiny because decisions made by law enforcement can have serious, potentially life-threatening impacts on citizens. Despite its established accuracy on certain training and testing datasets, facial recognition technology can falsely identify faces in the real world. The American Civil Liberties Union tested Amazon's Rekognition software on members of Congress, and the program misidentified 28 congresspeople as offenders.[51] The technology misidentified people of color at a disproportionate rate: "Nearly 40 percent of Rekognition's false matches in [the] test were of people of color, even though they make up only 20 percent of Congress."[52] Moreover, the ACLU's test cost only $12.33 using publicly available software marketed by Amazon, meaning that many people and organizations could cheaply access this flawed algorithm. Overwhelming public outcry has followed: Numerous stakeholders, from Amazon's employees to the ACLU, have made demands that Amazon halt its licensing of Rekognition to law enforcement. Using faulty facial recognition to identify criminals could lead to police misidentifying innocent individuals as possibly-violent suspects, leading to hostile confrontations that could have a fatal price. Moreover, the technology could exacerbate existing biases and discrimination by law enforcement, given its error rate among minorities.

Facial Recognition technology is already widely deployed by police departments across the United States. A 2016 study by the Center on Privacy & Technology at Georgetown Law found that in 16 states the FBI can compare suspects with driver's license photos without additional consent and that 117 million adult Americans are included in law enforcement facial recognition databases.[53] Transparency varies between implementations, and regulations are dangerously lacking, allowing for frequent searches across databases containing citizens who were not implicated in any crimes. This uncontrolled implementation creates several vulnerabilities when the algorithms are inaccurate. First, few police organizations check the veracity of identification systems before adopting them, and human oversight can be limited, because "without specialized training, human users make the wrong decision about half the time."[54] Moreover, possible challenges to free speech may arise when law enforcement monitors participants in political and civil rights protest, as has been done historically with lesser forms of surveillance. While this may be hard to imagine in the United States, in countries with more oppressive governments, facial recognition could be used to identify citizens who have participated in protests and apprehend them while they are moving through any public space. Facial recognition's accuracy and rapid, automated use raises fresh challenges in the domain of free speech and liberty.

---

[51] Jacob Snow, "Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots," *ACLU*, last modified July 26, 2018, https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28.

[52] Ibid.

[53] Clare Garvie Alvaro Bedoya and Jonathan Frankle, "The Perpetual Line-up," *Center on Privacy & Technology at Georgetown Law*, last modified October 18, 2016, https://www.perpetuallineup.org/.

[54] Ibid.

The concerns in law enforcement are exacerbated by facial recognition's history of mis-identifying non-male and non-white populations. Facial recognition technology has varying performance across gender and race. The MIT Media Lab completed a study into the performance of commercial facial recognition technology. It found dramatic disparities in accuracy when classifying male and female faces.[55] Datasets often lack non-white, non-male faces, and they under represent people of African descent. The "Labeled Faces in the Wild" database used to train and evaluate many facial recognition algorithms consists of 83.5% white and 77.5% male faces.[56] The study built a new dataset, the Pilot Parliaments Benchmark (PPB) to achieve balanced database with representation of various skin colors and sexes. Photos in the PPB were classified according to skin tone and sex; then, the several commercial facial recognition products were tested on the PPB. Each of the algorithms, including those produced by IBM, Microsoft, and Face++, performed better on lighter faces than darker faces, and on male faces than female faces. Darker-skinned females were misclassified most often, at rates of up to 34.7%, while lighter-skinned males were misclassified at a maximum rate of only 0.8%. These statistics point to severe shortcomings in current facial recognition algorithms, and they question the readiness of the algorithms' implementation in real world law enforcement scenarios.

The complexities surrounding facial recognition's implementation in law enforcement demonstrate many concerns that persist in other domains. Walmart's in store implementation of FaceFirst to identify potential shoplifters and prevent theft may produce disproportionate error rates that lead to concerns of discrimination, targeting darker or female individuals due to false positives or biased intent. Also, facial recognition can be used without consent of shoppers, just as law enforcement facial recognition can scan databases of citizens who have no prior violations. Furthermore, consider RealNetworks deployment of facial recognition in schools. The ACLU and Legal Defense Fund have argued against deployment in New York schools contending that "increased surveillance of kids might amplify existing biases against students of color, who may already be over-policed at home and in school."[57]Administrators and these organizations are concerned that when implementations extend beyond simply using facial recognition to control access to a campus and begin to influence how students are treated, racial bias will arise in these algorithms that have yet to ensure sufficient testing.

Even before a facial recognition system can be deployed, faces for the database must be collected. This leads to issues of consent and privacy. Many people's faces have been documented and added to databases without permission.[58] In January of 2019, IBM released a dataset of almost a million photos, all scraped from the photo-sharing website Flickr. It compiled them into a database; however, photographers and subjects were not notified. The database is now public. Photos can be removed but not without difficulty: "IBM requires photographers to email links to photos they want removed, but the company has not publicly shared the list of

---

[55] Joy Buolamwini and Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of Machine Learning Research*, 81, (2018): 1-15, accessed April 20, 2019, http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

[56] Ibid.

[57] Issie Lapowsky, "Schools Can Now Get".

[58] Olivia Solon, "Facial recognition's 'dirty little secret': Millions of online photos scraped without consent," *NBC Universal*, last modified March 17, 2019, https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921.

Flickr users and photos included in the dataset, so there is no easy way of finding out whose photos are included."[59] This single example displays the complexity of compiling an ethical database: Is the public sharing of photos online the same as consent to participate in a facial recognition dataset? How can consent be properly assessed, especially in cases like this where participants were given no initial choice? Moreover, given the persistence of hacking of company databases in recent years, databases that are supposedly secure may in fact be prone to hacking if they are not properly stored and encrypted.[60] These incidents could release people's private faces and faceprints to hackers and the public.

Governmental and private use of facial recognition can lead to ethical issues in privacy and liberty even if (and because) algorithms are accurate and unbiased. A recent study by the New York Times demonstrates how facial recognition can be shockingly invasive when used concurrently with pre-existing surveillance technology.[61] New York City's Bryant Park has video cameras that are live-streamed online to allow citizens to check park conditions. The New York Times utilized Amazon's facial recognition software to analyze the Bryant Park video stream to identify the faces of workers. They first fed the software public images of workers from companies in the area, then analyzed the video stream. Their system identified several individuals, including a local professor from SUNY College of Optometry, walking through the park. The software is publicly available and perfectly legal. Any entity in the U.S. could have run the test for only $60. Millions of cameras deployed by government, individuals, and companies exist in the U.S., and their conversion to a powerful facial recognition system is only one step away, as demonstrated by the study. The information gained from monitoring could be almost endless, including knowing when someone leaves for work, the friends and partners with whom they associate. Facial recognition has not yet been implemented on a national scale in this way; however, it is crucial to understand the ethical implementations of such a capacity and to restrict facial recognition's use, if so decided, before facial recognition is deployed on such scale.

Though public and commercial uses for facial recognition technology are raising many questions in the United States, China's implementation has already reached an alarming extent. The government has partnered with tech firms to implement widespread surveillance in public spaces, including "approximately one camera for every seven citizens."[62] The system has allowed for decreased crime but also governmental monitoring of citizens' movements, invading individuals' privacy and liberty. Police officers wear AI-enhanced vision technology, and the government has partnered with private companies, such as SenseTime and Megvii to put up millions of cameras enhanced with facial recognition technology.[63] Since 2016, the cameras have helped catch more than 2000 suspects; however, that likely does not justify for the government's unrestricted ability "to identify any of its 1.4 billion citizens within a matter of seconds" or "the

---

[59] Ibid.

[60] Aisha Al-Muslim Dustin Volz and Kimberly Chin, "Marriott Says Starwood Data Breach Affects Up to 500 Million People," *The Wall Street Journal*, last modified November 30, 2018, https://www.wsj.com/articles/marriott-says-up-to-500-million-affected-by-starwood-breach-1543587121.

[61] Sahil Chinoy, "We Built an 'Unbelievable' (but legal) Facial Recognition Machine," *The New York Times*, last modified April 16, 2019, https://www.nytimes.com/interactive/2019/04/16/opinion/facial-recognition-new-york-city.html.

[62] Ibid.

[63] Eamon Barrett, "In China, Facial Recognition Tech Is Watching You," *Fortune*, last modified October 28, 2018, http://fortune.com/2018/10/28/in-china-facial-recognition-tech-is-watching-you/.

ability to record an individual's behavior to predict who might become a threat."[64] Moreover, Human Rights Watch researcher Maya Wang commented, "The intention of these systems is to weave a tighter net of social control that makes it harder for people to plan action or push the government to reform."[65] Such applications of facial recognition can limit citizens' ability to organize and protest the government. The technology presents an alarming demonstration of the dangers of public of hyper-surveillance created by algorithms that can identify faces in mere seconds. Like other powerful technologies, it can be implemented for both good and bad ends, depending on the user's intent. Without restrictions of how facial recognition is used, more situations where ethical compliance is cloudy or ethics are even directly violated will arise in the United States and around the globe. Thus, a comprehensive ethical framework is necessary to address these various concerns.

An Ethical Framework for Facial Recognition

Facial recognition industry players have recognized dangers of the technology and significant ethical concerns. After calling for regulations in July 2018, Microsoft's president Brad Smith published a December 6th, 2018 letter in which he detailed Microsoft's own proposed ethical framework.[66] Smith asserts that legislation should counteract bias and discrimination by: 1. Requiring transparency; 2. Enabling third-party testing and comparison; 3. Ensuring meaningful human review; and 4. Avoiding use for unlawful discrimination. To provide for privacy, he argues that facial recognition systems must: 1. Ensure notice and 2. Clarify consent. Finally, to protect the integrity of democratic political systems and free speech, he argues for "limiting ongoing government surveillance of specified individuals," meaning that [67]

While these principles provide a good basis for ethical facial recognition, I argue that they are not sufficient to protect consumers and the public. Microsoft's proposals fall short in two important ways. First, the provision to protect against discrimination does not adequately ensure that accurate algorithms are employed. While enabling testing and human review provide good checks in the system, there is no defined threshold for accuracy of an algorithm and no mandate to check a facial recognition algorithm's accuracy when used to identify faces from different demographic groups. As the ACLU has presented, these additional restrictions are essential to ensure that discriminatory systems are not implemented. To preempt such systems, aggressive testing and approval are necessary. Second, Microsoft's suggestions regarding privacy do not provide enough consumer control. Some databases that have been around for years are being repurposed for the training of facial recognition algorithms. In these cases in which consent was not already provided, there is still no way for people to opt out. Further definitions need to be made of which photos and acts are public and therefore are allowed to be analyzed with facial recognition, and which are not.

Legislation and the court system in the U.S. and internationally have already begun to refine the legal and ethical principles that govern facial recognition. In an article published in 1999, Christopher Milligan argues that the then-young facial recognition and surveillance systems being incorporated into law enforcement arsenals are "inordinately intrusive into

[64] Ibid.
[65] Ibid.
[66] Brad Smith, "Facial recognition: It's time".
[67] Ibid.

individual privacy to such an extent that they chill personal autonomy.[68]" Moreover, it questions whether surveillance and facial recognition combine to violate the fourth amendment. The article contends that, though it does not violate established legal rights to privacy, the technology is still important to ethically control. This conclusion suggests that additional recognition on the part of governmental bodies would be required for meaningful protection of consumer and public privacy.

Legislation on both the United States and global levels has begun to tackle this issue. The Commercial Facial Recognition Privacy Act, legislation proposed in congress in 2019, would increase the informational burden of companies using facial recognition, and it would add consumer controls to how data is used.[69] Largely modeled off Microsoft's proposals, the legislation would require companies to further notify the public when facial recognition technology is being implemented and to restrict data's use.[70] On the international level, the EU General Data Privacy Regulation adds consumer rights to knowledge of data capture, the need for consent, and other model concepts that could prove useful in the U.S.[71] Both of these regulations highlight an increased burden of consent actualized by better notification of consumers and citizens as to the data being collected. In facial recognition applications, this can look like the provision of signage in public spaces or commercial properties where the technology is implemented. Moreover, the regulations begin to establish standards for data's usage and storage.

More than just legislation, ongoing litigation in Illinois has served as a challenge to limit facial recognition. The Biometric Information Privacy Act was passed in Illinois in 2008, making Illinois one of the most progressive states in limiting biometric (and thus facial recognition) data gathering.[72] The Act necessitates that any company seeking to gather and use facial recognition data must obtain written consent and not profit from someone's biometric data, including facial structures. Interestingly, recent court cases have put the law and facial recognition to the test: Google was sued for Google Photos' usage of facial recognition without express consent. The judge concluded that Google Photos did use facial recognition to scan users' faces without proper consent but that the technology's use did not harm consumers in a significant way and thus did not violate the law.[73] This ruling contrasts with a 2016 case testing Facebook's facial recognition when identifying people to be tagged in photos.[74] Why would two similar cases with very similar have such different outcomes? The answer lies in how Facebook and Google differ

---

[68] Christopher S. Milligan, "Facial Recognition Technology, Video," 299.

[69] Emily Birnbaum, "Senators introduce bill to regulate facial recognition technology," *The Hill*, last modified March 14, 2019, https://thehill.com/policy/technology/434166-bipartisan-senators-introduce-bill-to-regulate-facial-recognition.

[70] Taylor Hatmaker, "Bipartisan bill proposes oversight for commercial facial recognition," *TechCrunch*, accessed April 20, 2019, https://techcrunch.com/2019/03/14/facial-recognition-bill-commercial-facial-recognition-privacy-act/.

[71] "GDPR Key Changes," *EUGDPR.org*, accessed April 20, 2019, https://eugdpr.org/the-regulation/.

[72] "Biometric Information Privacy Act," 740 ILCS 14/, accessed April 20, 2019, http://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=3004&ChapterID=57.

[73] Jeff John Roberts, "Google, Facebook, and the Legal Mess Over Face Scanning," *Fortune*, last modified January 4, 2019, http://fortune.com/2019/01/04/google-face-scanning-illinois/.

[74] Jeff John Roberts, "Defeat for Facebook in Court Is Bad News for Firms That Scan Faces," *Fortune*, last modified May 6, 2016, http://fortune.com/2016/05/06/facebook-biometrics/.

in their use of facial recognition data. Google Photos merely groups like photos together for the convenience of users; when I go onto my Google Photos app and look at my own photos, the app shows me which photos contain pictures of my brother or any other person I have previously identified on my account. In contrast, Facebook can identify users' faces to third parties, such as friends of friends. Regulation of facial recognition will likely continue for decades to come.

While a complex topic, the ethics of facial recognition will continue to evolve as time goes on. Other technologies, from Internet of Things, which will connect everyday objects with the internet, to other biometric data, like fingerprint and eye scanners, will all concurrently push the definitions of privacy and the acceptable tradeoffs between efficiency and freedom. Facial recognition has immense capabilities, due to its ability to tap into artificial intelligence's greatest strengths and its scalability, whereby only a few cameras on a few street corners could allow a company or government to gain significant insight into citizens' lives. Because of this potential, even greater caution is necessary. Large-scale implementation should be preceded by extensive testing to verify whether the algorithms have accurate, fair results across gender and skin color. Moreover, the datasets used to train facial recognition algorithms and the faces analyzed by the technology everyday deserve privacy protections to assure proper use of something so personal yet readily visible as a face. Most likely, not every ethical question presented in this chapter will be addressed as quickly as facial recognition continues to develop. But, rather than despairing at society's eventual descent into an Orwellian world (as many facial recognition commentators indulge), I prefer to view the continued evolution of facial recognition as an exciting opportunity to watch the applications and ethics of one of the most important future technologies change over the course of a lifetime.

Works Cited

Al-Muslim, Aisha, Dustin Volz, and Kimberly Chin. "Marriott Says Starwood Data Breach Affects Up to 500 Million People." *The Wall Street Journal*. Last modified November 30, 2018. https://www.wsj.com/articles/marriott-says-up-to-500-million-affected-by-starwood-breach-1543587121.

Barrett, Eamon. "In China, Facial Recognition Tech Is Watching You." *Fortune*. Last modified October 28, 2018. http://fortune.com/2018/10/28/in-china-facial-recognition-tech-is-watching-you/.

"Biometric Information Privacy Act." 740 ILCS 14/. Accessed April 20, 2019. http://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=3004&ChapterID=57.

Birnbaum, Emily. "Senators introduce bill to regulate facial recognition technology." *The Hill*. Last modified March 14, 2019. https://thehill.com/policy/technology/434166-bipartisan-senators-introduce-bill-to-regulate-facial-recognition.

Buolamwini, Joy and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of Machine Learning Research* 81, (2018): 1-15. Accessed April 20, 2019. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

Chinoy, Sahil. "We Built an 'Unbelievable' (but legal) Facial Recognition Machine." *The New York Times.* Last modified April 16, 2019. https://www.nytimes.com/interactive/2019/04/16/opinion/facial-recognition-new-york-city.html.

Costello, Tom and Ethan Sacks, "New facial recognition tech catches first impostor at D.C. airport." *NBC Universal.* Last modified August 23, 2018. https://www.nbcnews.com/news/us-news/new-facial-recognition-tech-catches-first-impostor-d-c-airport-n903236.

Culbertson, Anthony. "Indian Police Trace 3,000 Missing Children in Just Four Days Using Facial Recognition Technology." *Independent*. Last modified April 24, 2018. https://www.independent.co.uk/life-style/gadgets-and-tech/news/india-police-missing-children-facial-recognition-tech-trace-find-reunite-a8320406.html.

Garvie, Clare, Alvaro Bedoya, and Jonathan Frankle. "The Perpetual Line-up." *Center on Privacy & Technology at Georgetown Law.* Last modified October 18, 2016. https://www.perpetuallineup.org/.

"GDPR Key Changes." *EUGDPR.org.* Accessed April 20, 2019. https://eugdpr.org/the-regulation/.

Hatmaker, Taylor. “Bipartisan bill proposes oversight for commercial facial recognition.” *TechCrunch.* Accessed April 20, 2019. https://techcrunch.com/2019/03/14/facial-recognition-bill-commercial-facial-recognition-privacy-act/.

Kruszka, Paul et al. “22q11.2 deletion syndrome in diverse populations.” *American Journal of Medical Genetics* 173, no. 4 (2017): 879-888. Accessed April 20, 2019. https://doi.org/10.1002/ajmg.a.38199.

Kuflinski, Yaroslav. “How Facial Recognition Works.” *Iflexion*. Last modified July 17, 2018. https://www.iflexion.com/blog/face-recognition-algorithms-work/.

“Labeled Faces in the Wild Home.” *University of Massachusetts – Amherst*. Last modified January 1, 2018. http://vis-www.cs.umass.edu/lfw/.

Lapowsky, Issie. “Schools Can Now Get Facial Recognition Tech for Free. Should They?” *Wired*. Last modified July 17, 2018. https://www.wired.com/story/realnetworks-facial-recognition-technology-schools/.

Menn, Joseph. “Microsoft turned down facial-recognition sales on human rights concerns.” *Reuters*. Last modified April 16, 2019. https://www.reuters.com/article/us-microsoft-ai/microsoft-turned-down-facial-recognition-sales-on-human-rights-concerns-idUSKCN1RS2FV.

Milligan, Christopher S. “Facial Recognition Technology, Video Surveillance, and Privacy.” *Southern California Interdisciplinary Law Journal* 1, no. 9 (1999): 295-333. accessed April 20, 2019. https://heinonline.org/HOL/Page?handle=hein.journals/scid9&div=7&g_sent=1&casa_token=&collection=journals.

Roberts, Jeff John. “The Business of Your Face.” *Fortune*. Last modified March 27, 2019. http://fortune.com/longform/facial-recognition/.

Roberts, Jeff John. “Defeat for Facebook in Court Is Bad News for Firms That Scan Faces.” *Fortune*. Last modified May 6, 2016. http://fortune.com/2016/05/06/facebook-biometrics/.

Roberts, Jeff John. “Google, Facebook, and the Legal Mess Over Face Scanning.” *Fortune*. Last modified January 4, 2019. http://fortune.com/2019/01/04/google-face-scanning-illinois/.

Roberts, Jeff John. “Walmart’s Use of Sci-fi Tech To Spot Shoplifters Raises Privacy Questions.” *Fortune*. Last modified November 9, 2015. http://fortune.com/2015/11/09/wal-mart-facial-recognition/.

Simonite, Tom. “Amazon Joins Microsoft’s Call for Rules on Facial Recognition.” *Wired*. Last modified February 7, 2019. https://www.wired.com/story/amazon-joins-microsofts-call-rules-facial-recognition/.

Smith, Brad. “Facial recognition technology: The need for public regulation and corporate responsibility.” *Microsoft*. Last modified July 13, 2018. https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/.

Smith, Brad. “Facial recognition: It’s time for action.” *Microsoft*. Last modified December 6, 2018. https://blogs.microsoft.com/on-the-issues/2018/12/06/facial-recognition-its-time-for-action/.

Snow, Jacob. “Amazon’s Face Recognition Falsely Matched 28 Members of Congress With Mugshots.” *ACLU*. Last modified July 26, 2018. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28.

Solon, Olivia. “Facial recognition’s ‘dirty little secret’: Millions of online photos scraped without consent.” *NBC Universal*. Last modified March 17, 2019. https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921.

Sutton, Mark. “Facial Recognition Technology: 10 buzzwords demystified.” *ITP.net*. Last modified November 23, 2018. http://www.itp.net/618391-facial-recognition-technology-10-buzzwords-demystified.

“The top 7 trends for facial recognition in 2019.” *Gemalto*. Accessed April 20, 2019. https://www.gemalto.com/govt/biometrics/facial-recognition.

West, Jesse Davis. “21 Amazing Uses for Face Recognition – Facial Recognition Use Cases.” *FaceFirst*. Last modified May 2, 2018. https://www.facefirst.com/blog/amazing-uses-for-face-recognition-facial-recognition-use-cases/.

AI and an Algorithmic Approach to Combat Gerrymandering

Onuoha Odim

Introduction

Redistricting has recently come into national spotlight because of increased awareness of the implication of district lines in representational power.[75] The shape of U.S. congress, state house, and state senate districts directly influences who has the potential to be elected to local office and, by extension, who has the power to create legislation that governs the lives of the people who live under their jurisdiction. With the advance of computers, semi-automated and automated redistricting is used more often to draw districts. Allowing computers to take into account every parameter that goes into drawing districts allows districts to be defined in a way that allows for more equitable representation, but it can also lead to more clandestine forms of vote dilution through sophisticated districting algorithms. This chapter discusses the implications of North Carolina's current partisan gerrymander court case and explores some algorithmic approaches taken by redistricting councils to re-draw partisan and racial gerrymandered districts. We will also discuss some limitations that are associated with trying to quantify and implement some federal and state redistricting rules that are inherently unquantifiable. Finally we will end by exploring the potential evolution of artificial intelligence in the space of equal representation and redistricting. There will be a number of questions proposed in this chapter on redistricting and AI, including considerations of who even has the right to redraw districts. I will answer these questions, but there is no one size fits all answer to every question proposed. These are questions that people who are charged with redistricting consider when drawing their district lines. These are also questions pondered by the United States Supreme Court every time a redistricting court case is brought to their bench. These questions will allow for a better understanding of both the current implications of legislation as it relates to redistricting and the intersection between these regulations on redistricting and quantitative methods to implement these rules.

Current North Carolina Supreme Court Case

In January 2018, the national non-profit organization Common Cause sued North Carolina's legislature because Common Cause believed the Republican controlled legislature intentionally lead their map makers to draw the State's remedial congressional district map to allow an electoral advantage to Republicans.[76] This case is known as *Rucho v. Common Cause* and has made its way up to the Supreme Court along with some other gerrymandering court cases like *Gil v. Whitford*, T*urzai v. League of Women Voters* and *Benisek v. Lamone.*[77] In this chapter we will focus on *Rucho v. Common Cause* because it is the most pertinent to understand how gerrymandering works in a practical sense. Unlike how *Gil v. Whitford* deals with

[75] Sherman, M. (2019, March 26). High court questions courts' role in partisan redistricting. Retrieved from https://www.apnews.com/340d4a8846c44e8da001b71868f8195c

[76] Aurora-Temple-Barnes. (2019, April 22). Rucho v. Common Cause. Retrieved from https://www.scotusblog.com/case-files/cases/rucho-v-common-cause/

[77] Redistricting and the Supreme Court: The Most Significant ... (n.d.). Retrieved from http://www.ncsl.org/research/redistricting/redistricting-and-the-supreme-court-the-most-significant-cases.aspx

gerrymandering on Wisconsin's 99 assembly seats*, Rucho v. Common Cause* deals with North Carolina's 13 Congressional Districts.[78] In this court case, Common Cause, the plaintiffs, argue that North Carolina's congressional map violates the First Amendment because it dilutes Democrat votes by partitioning the State is such a way that allows Democrats less congressional districts than they should theoretically have. Common Cause also argues that the 'gerrymandered' districts violate the equal protection clause of the 14th Amendment because it strips away people's ability to vote for their representatives and therefore have equal protection under legislation that can be implemented under their representative's jurisdiction. The plaintiffs also argue that the districts deny residents of Article I Section 2 and 4 of the Constitution because elections are conducted in an unfair way. These sections prohibit any unfair conducts when it comes to elections. Because Common Cause argues that North Carolina's districts were drawn in such a way that gives an unfair advantage to one party, the outcome therefore forces elections to allow Republican candidates greater weight and a higher probability of being elected.[79]

Common Cause argues that North Carolina's 2016 congressional map, adopted by the State's legislature after their previous map was struck down, was an unconstitutionally partisan gerrymandered map because it was drawn to favor Republicans 10 to 3. On March 3, 2017, the case was consolidated with another Supreme Court case about the unconstitutional packing and cracking of North Carolina communities by Republicans called *League of Women Voters v. Rucho*. After the Supreme Court decided with the defendants to halt the legal proceeding of their case and send it to the district court level, the three panel district court decided in favor of the plaintiffs in their claims: the 14th Amendment Equal Protection Clause, The First Amendment, and Article I of the Constitution.[80] On August 31st, the defendants then argued to stay the ruling pending Supreme Court review and the district court granted the motion.[81]

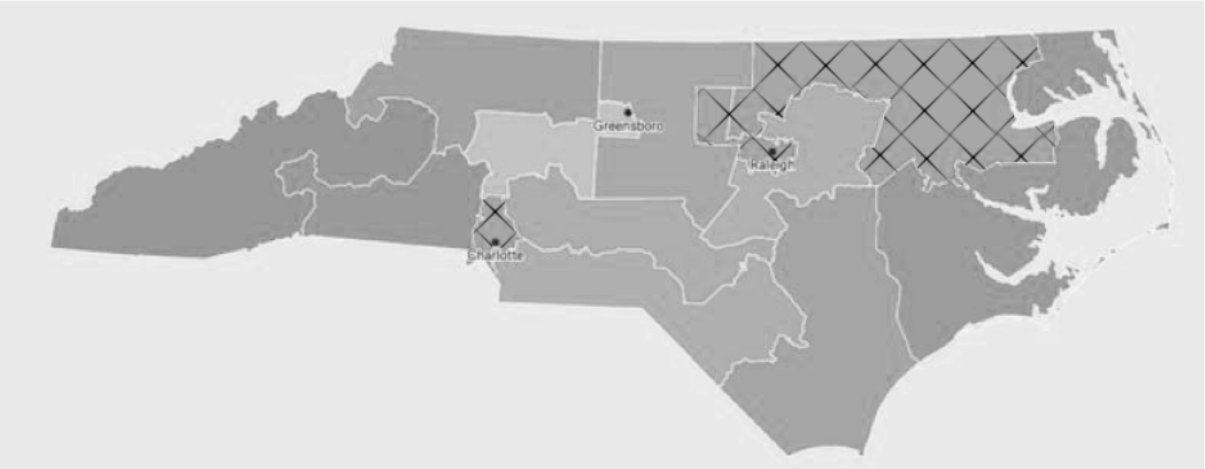

Figure 1. North Carolina Congressional District map. This figure shows the 13 Congressional Districts in North Carolina. Republican legislatures presides over 10 of the district while a

[78] Aurora-Temple-Barnes.
[79] Rucho v. Common Cause. (n.d.). Retrieved from https://www.commoncause.org/redistricting-litigation/common-cause-v-rucho/
[80] Ibid.
[81] Aurora-Temple-Barnes.

Democrat representative preside over 3 of the districts (shown crosshatched). Common Cause argues that this is an unfair distribution of electoral power because even though North Carolina voters vote half Democrat half Republican, only three out of the twelve congressional seats is held by a Democrat.
Map from *INDY Week*

During oral arguments on March 26th, 2019, Duke Professors Jonathan Mattingly and Gregory Herschlag used computer algorithms to demonstrate that North Carolina's remedial map was drawn in a way that allows democrats the most algorithmically minimum winnable districts possible.[82] When politicians manipulate voting maps like in the case of *Common Cause v. Rucho*(According to the three-judge panel finding the defendants to have unconstitutionally drawn the congressional map for partisan advantage),[83] they take from voters their ability to fairly and independently elect their own representative. To create cases like this where a party provides themselves a significant electoral advantage by portioning populations to algorithmically find the best possible district lines for their party, more sophisticated methods of artificial intelligence in population sorting is used to both bend the rules of redistricting to create partisan advantage and to meet congressional and state rules on redistricting.

Redistricting and Apportionment

Redistricting is the act of drawing district lines, be them Congressional, state House or state Senate district in the United States or legislative district maps in other countries. Every ten years after the U.S. census, districts have to be re-drawn to account for the shifting population in an act called apportionment. States, like Texas, have their state legislatures draw their respective districts.[84] Depending on the interests of the parties drawing lines, allowing legislatures autonomy in drawing districts can lead to districts that disenfranchise a minority race or a party's vote potential. Some states, like California, have independent redistricting commissions made up of a select group of bipartisan voters to draw the district boundaries that define the land area where each representative presides.[85] Commissions like these take the power of redistricting out of the hands of legislatures and proponents of independent commissions argue they lead to less partisan gerrymander districts.[86] Prior to 1965, gerrymandering was defined by the intent of legislatures to draw districts that disenfranchised some group's vote. This meant that to identify a district as gerrymandered one would first have to prove the person who created the district purposely drew it to allow an advantage to a particular group. Gerrymandering being reliant on intent rather than affect means gerrymandering is subject to conflicting interpretation of the rules that govern redistricting. Legislatures tasked with re-drawing districts might argue that they

---

82 https://today.duke.edu/2019/03/duke-mathematics-has-its-day-court?fbclid=IwAR0arDWbsygZKo1ooODI3Fa3f0zYoG8fu-GErpNtWGK7P6iuxWpYdArli84

83 Aurora-Temple-Barnes.

84 (n.d.). Retrieved from https://tlc.texas.gov/redist/requirements/congress.html

85 California, S. O. (n.d.). "Fair Representation - Democracy at Work!" Retrieved from https://wedrawthelines.ca.gov/

86 Soffen, K. (2015, July 01). Independently Drawn Districts Have Proved to Be More Competitive. Retrieved from https://www.nytimes.com/2015/07/02/upshot/independently-drawn-districts-have-proved-to-be-more-competitive.html

looked to ensure equal population first and compactness of the districts second or vice versa leading to an infinitely regressive argument of which rules should be considered above which and what order allows for the most equitable representational outcome. Through Section 5 of the Voting Rights Act of 1965, gerrymandering is now defined by the discriminatory effect of district shapes rather than intent by legislatures to be discriminatory.[87] Because proof of gerrymandering is then reliant on intent, algorithms and artificial intelligence can be used to measure the level of gerrymandering that exists in representative districts. In cases where districts are ordered by a court to be re-drawn, the districts have to be drawn to ensure proportional representation of the party or the minority demographic. In the case that the computer is sophisticated enough a re-drawing of gerrymandered districts, no matter how many iterations, can still lead to a partisan or racial gerrymander. Drawing districts to meet every state and federal redistricting rule, however, means developing a fully autonomous, computerized approach which would take in more parameters in the creation of representative districts.

AI and Algorithmic Approaches

The Voting Rights Act of 1965 prohibits any party from disenfranchising any racial or political group. The equal population federal redistricting requirement requires all representative districts to have equal (or near equal) population to ensure everyone's vote counts the same. It is a national requirement that Congressional districts comply with both the Voting Rights Act and the equal population redistricting requirement.[88] The order in which one rule is preferred over the other is subject to whatever the legislature drawing the district finds appropriate. In combination with the federal requirements for redistricting, all states also provide parameters that have to be met in drawing their state districts. Each state's congressional districts have to be compact with as few branching out of edges as possible. The closer a district looks to a circle the more compact it is because a circle is the only shape with a parameter that is dispersed perfectly evenly around its center.[89] Districts also have to comply with the state requirement of being contiguous which means no part of a district can be disconnected from the rest of the district. This rule allows districts to be closed off and the people under the legislative authority of the representative live in a similar geographical area.[90] A district like Illinois' 4th Congressional District which connects two predominantly minority neighborhoods through a high way and looks like a pair of earmuffs is technically contiguous but is not compact. States also require their districts to keep 'communities of interest' together. This means that like communities presumably have the same political interests so therefore should be kept together.

The communities of interest parameter is inherently hard to define or operationalize because there is no clear identifier of what constitutes a community of interest. A community of interest can be defined by the similar interest of a demographic population that lives in a particular area, or it can be defined by the political interest of a group that lives in a geographically unique area; like individuals living near a community lake who are interested in the protection of the lake.[91] States also look to ensure their districts are fair and competitive.

---

[87] Redistricting Criteria: The Voting Rights Act. (n.d.). Retrieved from http://www.publicmapping.org/what-is-redistricting/redistricting-criteria-the-voting-rights-act

[88] (n.d.). Retrieved from http://aceproject.org/ace-en/topics/bd/bdb/bdb05/

[89] Wendy.

[90] Ibid.

[91] Ibid.

Fairness in districts is necessary so any political party can have the same ability to win a district, but this redistricting rule is sometimes complicated to operationalize. Ensuring fairness necessarily means the person who is drawing the districts has to have access to past voting preferences of the people living in the area. This access to voting data allows the individual to create some bias in the creation of the districts. A fully autonomous redistricting computer can use prior voting trends to create the most functionally fair districts possible but taking into account the rest of the national and state redistricting rules in the creation of districts makes it particularly complicated for autonomous bots to draw districts without at least some human assistance.

A redistricting rule that is particularly hard to operationalize autonomously is partitioning districts in a way that coincides with major streams or roadways to ensure houses are not split down the middle in creating districts. To draw the best line that splits a county by a major roadway it is necessary for someone to identify and tell the bot which roadway should be considered over which. The alternative is to use GIS to re-create the identified roadway and satellite imaging to autonomously decide which roadway should be partitioned over which. The potential computational power this operation requires is presumably too high considering there are hundreds of thousands of VTDs and census blocks the GIS would have to scan through. There are more rules governing state senate and house redistricting than congressional districts because state districts are subject to both national and state redistricting rules while congressional districts are subject only to national rules. Intelligent districting mechanisms would take into account each federal and state requirement for redistricting. This intelligent mechanism would mean keeping track of all the parameters for redistricting including compactness, contiguity, equal population, compliance with the VRA, preservation of existing political communities, partisan fairness and racial fairness.

Philip Klein, a computer scientist at Brown, is leading an algorithmic approach to redistricting that solves for partisan bias in district plans.[92] His team uses a method for dividing populations into compact districts with six or fewer sides, which is a huge contrast from current gerrymandered districts that come in all sorts of shapes. Professor Klein explains his team's algorithmic approach is an attempt to take the partisan bias out of redistricting. The approach uses a clustering method called 'K-means' clustering which is a resource allocation algorithm that is used in determining the distribution of limited resources. For example, the efficient distribution of fire stations throughout a city or the efficient allocation of stores to maximize convenience for customers can be identified using a K-means algorithm. In the case of a congressional district, the resource that would be allocated is the number of congressional seats in a state. The efficient allocation would be the number of citizens in that state divided by the number of congressional seats so each person's vote counts the same. Using a Voronoi diagram the collection of voters can then be coordinated off in discrete geometric clusters. These diagrams are convex and compact allowing them to be great representations of what districts should look like. Partitioning of voters into these clusters with equal populations that correspond to one congressional seat then allows for the national and state redistricting rule of 'equal population' to be met.

---

[92] Cohen-Addad, V., Klein, P. N., & Young, N. E. (2018). Balanced centroidal power diagrams for redistricting. *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL 18*. doi:10.1145/3274895.3274979

Brian Olson a software engineer in Massachusetts created an algorithm that engineers districts without any bias.[93] His algorithm draws compact districts to have equal population while also respecting political boundaries like counties and cities. Compared to current congressional districts, Mr. Olson's computer drawn districts are more compact and create less arbitrary partitions. Since algorithms like his prioritize compactness, because it's a federal districting requirement, and allows voters who are geographically closest to be in the same electoral region it ensures coherence in districts. The problem with Brian Olson's approach is that it is unclear whether the districts drawn allow for equal representation potential for both democrats and republicans. However, According to Common Cause, North Carolina residents are split nearly evenly for their party preferences, meaning Olson's map might allow more proportional representation. Olson's method might then be effective in addressing North Carolina's partisan gerrymandering problem.

Figure 2. The first North Carolina district map is the map that Brian Olson made using his algorithm. The second map is North Carolina's current congressional district map. Olson's method takes into account the average distance between voters and the center of their districts while the current map does not take into account distance because it was drawn by state legislatures. Mr. Olson's method has a population deviation of 431 people while North Carolina's current district map has a population deviation of 1 person. This means that Olson's districts are not equally proportional.
Map from Brian Olson's website *bdistricting.com*

Depending on the advancement of geographic information systems, drawing congressional and state districts with no human interference will be the next step of redistricting.[94] Artificially limiting the extent to which humans can interfere with districts created by a computer is the first step to complete autonomy of geographic information systems (GIS)

[93] Redistrict. (2018, January 25). The Atlas Of Redistricting. Retrieved from https://projects.fivethirtyeight.com/redistricting-maps/
[94] Smith, R. Duke Mathematics Has Its Day in Court. (n.d.). Retrieved from https://today.duke.edu/2019/03/duke-mathematics-has-its-day-court?fbclid=IwAR0arDWbsygZKo1ooODI3Fa3f0zYoG8fu-GErpNtWGK7P6iuxWpYdArli84

redistricting tools.[95] Replacing human judgment with just a set of hierarchical rules for redistricting will mean biases that are inherent in human control of redistricting will go away. Since the VRA calls for proportional representation, outcome where there is a 40, 60 split in a state for two parties would mean that the electorate has to be representational of the voters meaning 40% of the districts will belong to the minority party. It is already known that allowing individuals the full autonomy to partition populations with the intent to allow political power to a few can be lead to district manipulation and possibly provide the advantage to a single party. Considering a legislature's advisory redistricting council is able to take into account past election data and the location where incumbents live when creating districts, it is not shocking to see how districts can be manipulated to favor the legislators that are charged with re-drawing district lines. With computers becoming more and more efficient, is has become increasingly easier to use powerful computers to draw districts that gerrymander without the clear indication of gerrymandering.[96] Instead of drawing districts that are oddly shaped and partisan, advanced computers are now capable of drawing districts that take into account every compactness and contiguity score, therefore are able to draw districts that look geographically fair, while also being just as unfair as any partisan hand-drawn district.

There are two limitations to computer redistricting. The first is the inability to create a mathematical representation of optimal compact, contiguous, and equal population districts. This problem is known as an "NP-hard" partitioning problem.[97] Computationally, the chances of optimal contiguity decreases while at the point of optimal contiguity of a district the optimal proportional population decreases. This inability to perfectly optimize these redistricting rules means the addition of the others makes it nearly impossible to create a perfectly representational district map. The second limitation is quantifying qualitative redistricting requirements computationally like communities of interest. This conflicting optimum of redistricting rules creates an inability to create districts that are always proportionally representational. Translating redistricting criteria, like the Voting Rights Act, also creates a layer of complication that proves to be intractable when taking into consideration the hundreds of Voter Tabulation Districts that exist in any given state. Quantifying the levels of segregation that exists throughout a state and keeping like communities together while also taking into account how to partition neighborhoods in a way that allows the proportional number in a given state to align with the exact voting preferences of the state's population also adds to the computational intractability of redistricting.

Ethical Implications of AI

The advantage of using artificial intelligence with redistricting means district map makers can create districts that can potentially comply with federal and state requirements of districts but still allow an advantage for a party. This clandestine gerrymandering would mean that artificial intelligence can be used to more efficiently strip individuals from their ability to elect their preferred representative. But, without the use of artificial intelligence it would be nearly impossible to meet every single federal requirement for a district while also ensuring no party gets an unfair advantage in an election. Independent redistricting commissions are non-partisan

[95] Oberhaus, D. (2017, October 03). Algorithms Supercharged Gerrymandering. We Should Use Them to Fix it. Retrieved from https://motherboard.vice.com/en_us/article/7xkmag/gerrymandering-algorithms

[96] Perzanowski, A., & Schultz, J. (2016). The Promise and Perils of Digital Libraries. *The End of Ownership*. doi:10.7551/mitpress/9780262035019.003.0006

[97] Ibid.

bodies charged with drawing districts during apportionment that functionally meet federal and state redistricting requirement. These commissions might be able to navigate the potential subversive ability of one party who has the full autonomy and power to create districts in whatever shape they like but they are still represented by individuals that align with a particular party and therefore are subject to the same biases and party preferences that come with wanting your party to have an advantage in any electoral cycle. Artificial intelligence and autonomous redistricting bots on the other hand do not hold the same bias that is inherent in humans but their implementation creates a unique layer of complication in the creation of districts that still allows for biases to happen depending on the people in control of the algorithms. So, is it better to allow independent redistricting commissions draw districts? Artificial intelligence with specific algorithms designed to be as non-partisan as possible to draw districts? Or just keep allowing legislatures the ability to draw their own districts? There are a number of disadvantages and questions regarding complex algorithmic approaches to redistricting that I will answer here.

How do algorithms take into account keeping communities of interest together when drawing districts?

Algorithms take into account keeping communities of interest together by following the lead of the legislatures in charge of drawing the districts. These legislatures know their state best and understand how communities shift from neighborhood to neighborhood. Using local knowledge, legislatures are able to make the best guess as to where districts should be drawn that is as least obtrusive to communities. These legislatures might not walk neighborhood to neighborhood to understand the nuances that separates them, but their contemporary knowledge of their state together with their use historical political boundaries they input their best understanding of where local communities should be partitioned into two separate districts.

Is it possible to quantify a community of interest and should these community's boundaries lie separately from city and county boundaries?

There is no right way to partition a county to create districts but legislatures are charged with using their understanding of their state to create these partitions. Algorithms will necessarily need the help of legislatures to understand what qualitatively differentiates one community from another because these parameters can easily change year to year by new people coming into communities or just a shift and evolution in interest and mindset of the people who living in the area. Community of interest boundaries should correspond with city and county boundaries. But, when splitting a city through the center, it is necessary to find the best line that separates as few neighborhoods with similar political interests.

Should identifying and preserving communities of interest even be a concern considering there are more pertinent redistricting requirements that need to be met, like contiguity and compactness?

Identifying and keeping communities of interest as intact as possible when creating representative districts is necessary to allow the district to function like it should. If a community is intentionally split, separating people with similar political interests then neighborhoods will not be able to call for local change as effectively as they should be able to. Collective action for local problems becomes fragmented when people who are interested in solving those problems are separated. It is necessary to allow these communities with similar interests to stay together so

their political interests can be easily identified and addressed by their representative. Keeping communities of interest together however should not take precedence over national redistricting requirements like contiguity and compactness because it is necessary for a district to be as contiguous and compact as possible to ensure districts keep similar shapes, fit each state, and separate populations in the most efficient manner.

Does drawing lines to group voters who are closest together even ensure the coherence of communities of interest? What even is a community of interest?

A community of interest can be looked at as racially similar neighborhoods or parts of a city that are in the same school districts. These communities are tied together by an interest like preserving their representation in congress or ensuring their children receive the best education possible. Because of these political interests preserving the communities allows them to be more incentivized to act of political matters.[98]

The Voting Rights Act also mandates race be accounted for when drawing districts in some states and most algorithms do not do this. But, should minorities even be packed into one district to ensure their political representation?

Minorities should be allowed proportional representation. The intent of the Voting Rights Act was to ensure groups of people who have been historically disenfranchised are able to elect their preferred representative with no barriers or illegal opposition. Redistricting is the most legal way to suppress the political interest of a particular group, be it racial group or political party. In ensuring proportional representation it is necessary to draw districts in such a way that allows the chance for racial minorities to be represented in the legislature. If African Americans make up 30% of the state, and there are 20 congressional seats, it is equitable for African Americans to have the potential to elect at least 6 representatives of their choice to fill those 20 congressional seats. Having equitable representation like this allows the intent of the Voting Rights Act to be met.

How do algorithmic approaches preserve boundaries? Should counties necessarily be partitioned to allow equal population or should preserving county boundaries be a more important rule than the preservation of equal populations in districts?

Algorithmic approaches to drawing districts only take into account the parameters that were programmed into them. This means that their inherently has to be a human laying out the framework of how districts should look to ensure every national and state redistricting rule is met. In cases where a state's district is found to be gerrymandered algorithmic approaches is the most effective way to identify the level of gerrymandering. Having expert reporters who are familiar with racially polarized voting, redistricting requirements, and the algorithms behind districts is the most effective way identify if a district is gerrymandered. In North Carolina's case, Duke's expert reporters served as the people with the foremost knowledge on districting and through countless iterations of North Carolina's districts in computer software like QGIS, they were able to find significant evidence that North Carolina's districts were gerrymandered to

[98] Redistricting Criteria Underhill, Wendy (2019) Retrieved from http://www.ncsl.org/research/redistricting/redistricting-criteria.aspx

allow Republicans the statistically highest advantage they could have.[99] Even though the Supreme Court has not made a verdict on this case, expert reporters are tasked to provide strong cases and evidence for or against gerrymandering claims that leads to objective conclusions and this can only be provided through algorithms and statistical analysis of districts.

How do redistricting methods take into account the Voting Rights Act to protect against vote dilution of majority-minority communities, or should this not be a concern?

The order in which redistricting rules are taken into account is dependent on the weights given to rules in computer programming. These weights are not standardized but the most efficient weight that allows the rules to most effectively draw districts is preferred. Placing contiguity over compactness might shift the way districts are partitioned but identifying if that order provides for more equitable representation in combination with the other rules is what legislatures who are charged with redistricting look for. Attempting to ensure every redistricting rule is met however places a unique challenge for legislatures and this is where the complicated nature of redistricting lies. Gerrymandering tests where experts compare one map to thousands of iterations of that map preserves political boundaries by only shifting the edges of the districts rather than creating new districts. This means that political boundaries are kept at the same level only shifted slightly to include communities that were previously not in the district to see if that inclusion would lead to maps that more adequately represents the state's population. Redistricting methods use counties as the largest borders to keep together. It is first necessary to keep a county intact when drawing districts. If it is necessary to split a county in order to have districts with the same population, then it is necessary to find the best location to split those counties. Aside from keeping cities within counties as intact as possible it is necessary to keep alike communities intact too. This is where the question of communities of interest comes in. Redistricting should be as transparent as possible. Because districts are boundaries that represent what area a legislature governs, the people who are in their districts who elect that representative should have a right to know how their district was drawn to include or exclude some people. They should have the right to know the reasons for their district to be shaped the way it is so they can better understand the political interest of their communities and therefore come together more effectively.

Finally, how transparent should redistricting be? Should we, as voters, all get a say in who is in our districts? It seems like computational redistricting methods like Mr. Klein's and Mr. Olson's methods only really solve the problems of equal population and prevention of partisan bias but ultimately leave open a number of other problems and ethical concerns that are tied with redistricting.[100,101]

Evident in the North Carolina case, the effects district shapes have on who can even be elected is so strong that every Democrat in a state can vote for their respective Democrat candidates in their district while just some Republicans in the state can vote for the Republican

---

[99] Smith.

[100] Researchers devise an algorithm to combat gerrymandering. (n.d.). Retrieved from https://phys.org/news/2017-11-algorithm-combat-gerrymandering.html

[101] Cohen-Addad, V., Klein, P. N., & Young, N. E. (2018). Balanced centroidal power diagrams for redistricting. *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL 18*. doi:10.1145/3274895.3274979

candidate and still the Republicans might have a majority in the house because of the way the populations are partitioned.[102] The general population should be aware of this reality so they can understand how much of an impact their vote is actually having on local elections. Increased understanding of the redistricting process will also lead to representatives to have their minds on the wants of their constituents because the people who they are supposed to represent are more aware of the political process to its most minute level which is redistricting. Having an increased understanding of redistricting would lead to both a more informed citizenry and an increase in fiduciary responsibilities of representatives.

Future of AI and Redistricting

There is no universal agreement on what is fair. When it comes to redistricting, there are a number of complicated redistricting rules that necessarily makes it hard to operationalize every national and state redistricting requirement. In this sense it is necessary to prioritize rules like compactness, contiguity, equal population and compliance with the Voting Rights Act over any other redistricting rule. Currently redistricting is semi-autonomous; meaning a combination of GIS and human assistance creates congressional, state house and state senate districts. The future of redistricting lies in full autonomy.[103] As computers become more sophisticated, more redistricting parameters will be able to be taken used to create districts. This means not only will compactness, contiguity, equal population and compliance with VRA be used to create every single district, preservation of existing political communities, partisan fairness and racial fairness, coincidence with major roads and coincidence with census tracts will also be taken in the creation of districts.

The nature of redistricting is inherently complicated because of the number of rules that have to be implemented. Including the fact that some of these rules require translation for computers, like communities of interest and coincidence with major roads, it also adds to the computational barrier in drawing districts. But, because of increases in sophistication of computers, we will be able to code into them more parameters that allow increasingly complex models that can incorporate more information into the creation of district boundaries. The future of redistricting is almost certainly a combination of independent commissions and artificial intelligence. With the input of a bipartisan force that has the sole aim at making sure representational districts are drawn fairly and with the help of artificial intelligence that can effectively incorporate all the rules of redistricting, the bounds for gerrymandering can possibly tighten. Another reality is the potential for partisan committees to use algorithms and AI to supercharge gerrymandering. This clandestine gerrymandering means drawing districts too look like they comply with national and state redistricting rules while in fact they give an unfair advantage to a party. If a district is drawn in this way the best method is to use QGIS and other analytical software to compare the district with thousands of iterations of itself to see by just how much of an advantage is one party getting.

The question of gerrymandering is left to interpretation by the Supreme Court and interpretation can easily shift depending on the analysis provided by expert reporters who run the iterations of districts. If the experts find that a state's Congressional, House, or Senate district map is drawn to heavily favor a party, or disenfranchise a racial group, then arbitration in combination with precedence by either the US Supreme Court or by local courts dictate whether

[102] Smith.
[103] Ibid.

the district is gerrymandered and has to be redrawn.[104,105] Algorithms are used to both draw districts that take into account every federal and state redistricting requirement and test if current districts were drawn to be fair.[106,107]

[104] Hor, M. (2010). Intelligent electoral districting mechanism. *2010 International Conference on Machine Learning and Cybernetics*. doi:10.1109/icmlc.2010.5580893

[105] Duke Mathematics Has Its Day in Court. (n.d.). Retrieved from https://today.duke.edu/2019/03/duke-mathematics-has-its-day-court?fbclid=IwAR0arDWbsygZKo1ooODI3Fa3f0zYoG8fu-GErpNtWGK7P6iuxWpYdArli84

[106] Cohen-Addad, V., Klein, P. N., & Young, N. E. (2018). Balanced centroidal power diagrams for redistricting. *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL 18*. doi:10.1145/3274895.3274979

[107] Duke Mathematics Has Its Day in Court. (n.d.). Retrieved from https://today.duke.edu/2019/03/duke-mathematics-has-its-day-court?fbclid=IwAR0arDWbsygZKo1ooODI3Fa3f0zYoG8fu-GErpNtWGK7P6iuxWpYdArli84

Works Cited

Achen, C. H., & Shively, W. P. (1995). *Cross-level inference*. University of Chicago Press.

Aurora-Temple-Barnes. (2019, April 22). Rucho v. Common Cause. Retrieved from https://www.scotusblog.com/case-files/cases/rucho-v-common-cause/

California, S. O. (n.d.). "Fair Representation - Democracy at Work!" Retrieved from https://wedrawthelines.ca.gov/

Cohen-Addad, V., Klein, P. N., & Young, N. E. (2018). Balanced centroidal power diagrams for redistricting. *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL 18*. doi:10.1145/3274895.3274979

Gardner, J. A. (2001). One Person, One Vote and the Possibility of Political Community. NCL Rev., 80, 1237.

Goodman, L. A. (1959). Some alternatives to ecological correlation. *American Journal of Sociology*, *64*(6), 610-625.

Hor, M. (2010). Intelligent electoral districting mechanism. *2010 International Conference on Machine Learning and Cybernetics*. doi:10.1109/icmlc.2010.5580893

Ingraham, C. (2016, January 13). This is actually what America would look like without gerrymandering. Retrieved from https://www.washingtonpost.com/news/wonk/wp/2016/01/13/this-is-actually-what-america-would-look-like-without-gerrymandering/?noredirect=on&utm_term=.8b3d355ac4e3

King, G. (2013). *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data*. Princeton University Press.

(n.d.). Retrieved from https://tlc.texas.gov/redist/requirements/congress.html

Redistricting and the Supreme Court: The Most Significant ... (n.d.). Retrieved from http://www.ncsl.org/research/redistricting/redistricting-and-the-supreme-court-the-most-significant-cases.aspx

Redistricting Criteria Underhill, Wendy (2019) Retrieved from http://www.ncsl.org/research/redistricting/redistricting-criteria.aspx

Researchers devise an algorithm to combat gerrymandering. (n.d.). Retrieved from https://phys.org/news/2017-11-algorithm-combat-gerrymandering.html

Rucho v. Common Cause. (n.d.). Retrieved from https://www.commoncause.org/redistricting-litigation/common-cause-v-rucho/

Sherman, M. (2019, March 26). High court questions courts' role in partisan redistricting. Retrieved from https://www.apnews.com/340d4a8846c44e8da001b71868f8195c

Smith, R. Duke Mathematics Has Its Day in Court. (n.d.). Retrieved from https://today.duke.edu/2019/03/duke-mathematics-has-its-day-court?fbclid=IwAR0arDWbsygZKo1ooODI3Fa3f0zYoG8fu-GErpNtWGK7P6iuxWpYdArli84

Soffen, K. (2015, July 01). Independently Drawn Districts Have Proved to Be More Competitive. Retrieved from https://www.nytimes.com/2015/07/02/upshot/independently-drawn-districts-have-proved-to-be-more-competitive.html

I spy with my little eye? - Artificial Intelligence and Fourth Amendment Search

Analese Bridges

The Ethics of AI in Journalism

Jackie Park

Artificial intelligence has saturated nearly every professional field, and the journalism industry is no exception. The journalistic process of gathering, assessing, creating, and presenting news and information is no longer just a human endeavor, as many newsrooms around the world have started to utilize AI technology. The United Kingdom's Press Association collaborated with technologists at Urbs Media to create an AI machine that produces 30,000 localized news reports every month. Other major news outlets such as The Washington Post, The New York Times, and so on, have also followed this trend.[108]

But there are complications in the journalistic procedure that have in the past or could in the future arise from the implementation of AI in the journalism industry. In this chapter, we will review these potential issues with a focus on the ethical implications. We will refer to the field of journalism as the journalism industry and the media interchangeably, as the term "media" refers to news outlets that the public trusts and depends on for news and information. We will first review the ethical obligations of the media to the public, which comprise of truthfulness and public service. These categories will serve as a framework through which we appraise AI's role in the newsroom. The scope of research focuses on democratic media environments in which the media occupies a public role but operates independent – relative to non-democratic societies – from the government. We will perform a historical analysis on the current implementations of AI in newsrooms to evaluate its potential as well as shortcomings. This section details the dynamic between AI and human journalists in news outlets and what AI incorporation indicates in terms of economic efficiency and journalistic ethics. The time range for evaluation will be a decade; because AI implementation in the media became prevalent in the 21st century, the past decade will be our main source of evaluation and likewise, the upcoming decade will be our parameters of reference to the future. We will then address the ethical concerns pertaining to AI implementation in journalism, touching on topics such as declining trust in the media, privacy violations, and the negative impacts on human journalists. Lastly, we will make recommendations based on our analysis on ways newsrooms should go about utilizing AI going forward, beyond just 10 years.

Ethical Duty of the Media in Democratic Societies

The three traditional duties of the media in a liberal democracy can be summarized as the following: (1) monitoring authorities and public figures to prevent or expose incompetence or abuse of authority, (2) representing the voice of the public, and (3) providing basic knowledge on affairs that may affect the public's day-to-day lives such as information about an electorate.[109] Because regular citizens do not have the access or resources that the media is granted, they depend on the media to use their advantage in order to properly serve these duties. The trust

[108] James Ovenden, "Will AI Destroy Journalism? The Profession Is Struggling and AI Could Be the Final Nail in the Coffin," Innovation Enterprise, December 21, 2017.
[109] Carl Fox, "Public Reason, Objectivity, and Journalism in Liberal Democratic Societies," Res Publica 19, no. 3 (2013): , doi:10.1007/s11158-013-9226-6.

between the media and the people is elemental for a functional society. Only with this trust could people safely utilize news and information to determine the ways in which they go about their individual lives and approach issues beyond, whether it's the daily weather forecast or breaking news on a political scandal. There are multiple ethical obligations journalists must abide by in order to achieve these functions.

Truthfulness

Truth as a general concept is put on a pedestal, for the human search for truth is a natural desire attached to the need to know how the world works. But the necessity of truthfulness varies depending on the status of the speech.[110] The value of truthfulness declines when the exposure of it results in harming other societally valuable attributes such as children's innocence or basic human dignity.[111] Parents hide the truth about Santa Claus or the tooth fairy from their children without much moral conflict; on the contrary, those who expose somebody's personal information without any societally beneficial purpose are shunned, regardless of factuality. In the same manner, the value of truthfulness enhances when the absence of it can be detrimental to society. This is the reason that false information in selling goods, services, and other financial instruments is strictly regulated. It is also the reason that – with respect to the public's dependence on the media – truth is elemental in news.

The definition of truth has long been debated, but it is not necessary to define the philosophical details of truthfulness in order to arrive at workable guidelines for journalistic truth.[112] While journalistic truthfulness is easily determined for information that is factually black-or-white (it either happened or it didn't), there are some journalistic aspects that touch on a gray area. The interpretive dimension of journalism such as editorials or investigative reports is often labeled "false" for lacking objectivity – a common standard of measuring journalistic truth. However, it is impossible for human beings to completely eliminate personal bias. So to expect complete objectivity from human journalists is idealistic and impractical.[113] That is not to say journalists could be as prejudiced as they'd like. Rather, the standard for measuring journalistic truth must shift from objectivity to the *pursuit* of objectivity.

This shift is crucial considering the possibility of unintentional mistakes. To deem a news outlet a liar for an unintentional mistake or a hint of prejudice would be an unreasonable strain on the much needed trust between the media and the public. Thus, people must trust that the media constantly strives for truth, which they could measure by evaluating not the outcome but the process through which the journalists gathered information – whether they selected sources based on tangibility and credibility and whether they considered varying perspectives.[114] As long as there is evidence of a journalist's attempt to create a method to manage personal biases and external manipulation, their work still holds truthful.[115] And because of the intrinsic trust the

---

[110] Jackie Park, *Fake News as a Threat to the Democratic Media Environment*, Senior Thesis, Duke University, 2018.
[111] Frederick Schauer, "Facts and the First Amendment," UCLA Law Review 57, no. 4 (2014).
[112] Dale Jacquette, JOURNALISTIC ETHICS: Moral Responsibility in the Media (ROUTLEDGE, 2017).
[113] Fox, "Public Reason".
[114] Brent Cunningham, "Re-thinking Objectivity," Columbia Journalism Review 42, no. 2 (July/August 2003):, https://archives.cjr.org/feature/rethinking_objectivity.php.
[115] Stephen J. Berry, "Why Objectivity Still Matters," Nieman Reports 59, no. 2 (Summer 2015): , https://niemanreports.org/articles/why-objectivity-still-matters/.

public has towards the media in democratic societies, journalists have an ethical obligation to always strive for truthfulness.

AI manages the shortcomings of achieving truthfulness in ways that humans are incapable of. An error in truthfulness due to bias is not an issue for AI systems that do not develop personal opinions. An error in truthfulness due to lack of information is also less likely, as AI could gather and process a vast scope of data within seconds. But what if AI systems are programmed incorporating bias? If human beings, prone to prejudice, are the masterminds behind AI algorithms, is it possible to program without bias? And are there other ethical implications in utilizing data the way AI does regardless of its efficiency? These are questions that will be addressed in following sections that detail AI's role in newsrooms.

Public Service

Another journalistic ethical obligation stemming from the need to maintain the public's trust is to supply information that serves public interest, even when it conflicts with personal interest. We define public interest in alignment with the 2nd traditional duty of the media mentioned above: representing the people. To serve this purpose, the term "public interest" should be interpreted on two different levels – the general and the individual. The media could determine general public interest by keeping up with what is relevant in every hour of every day. For instance, a local station would report on crime assuming that the area's residents would be interested in such information for the sense of security. Economic and political occurrences such as stock market updates or policy changes that impact a great number – if not all – of the society's participants also meet the criterion for general public interest.

But the media must also attempt to represent and amplify each and every individual's voice. Due to the diversity of ideas existing in a society, the media must provide an environment for diverse public discourse. The beginning of the free discourse ideal could be traced to the Enlightenment in the 18th century. The underlying belief of this movement was that the abolishment of governmental censorship would allow free public discourse, which would spark intellectual and scientific discoveries that would contribute to the overall advancement of the society.[116] The marketplace of ideas theory, introduced in the 20th century, furthered this conviction.[117] Per Adam Smith's economic belief that a free market will prosper without regulations, this theory suggested that hands-free competition of unrestrained discourse will ultimately lead to the most favorable outcome: the triumph of the best quality information, followed by a well-informed public.[118] Although the question of whether more discourse leads to the best information is debatable, because liberal democratic societies today are enrooted in the Enlightenment and/or adhere to the power of Smith's theory, participants of these societies expect the media to help achieve this ideal. Under this definition of public interest, the media's ethical obligation of public service is not only to present stories that provide collective public service, but also those that any one person may be interested in, whether that be the corrupt actions of a political figure or the results of a local high school baseball match.

---

[116] Sue Curry. Jansen, Censorship: The Knot That Binds Power and Knowledge (New York: Oxford University Press, 2010).

[117] John Milton, Areopagitica, 1644 (New York: Published by Columbia University Press for the Facsimile Text Society, 1927).

[118] Adam Smith, The Wealth of Nations (New York: American Home Library, 1902).

As we segue into evaluating the ethical implications of AI implementation in the media, we must have these journalistic responsibilities in mind. We must appraise whether AI in journalism hinders the media's ability to meet these ethical standards. Now, we are not proposing that human journalists have successfully met these obligations. But we will assume that they nevertheless strive to achieve the mentioned goals as human beings with some degree of moral compasses. Whether AI could be programmed with the same moral pursuit will be further discussed throughout the chapter.

Robot Journalists vs. Human Journalists

The journalistic process can be broken down into 4 main steps: (1) assignment planning, (2) gathering content, (3) production, and (4) publication. Different skill sets are required to accomplish each of the four stages of the process and thus, AI implementation in each stage varies as well. In the first stage of assignment planning, AI makes topic selection more efficient by choosing topics that are newsworthy through big data and correlation analysis. Through measurable units such as likes, shares, and reposts on social media platforms, robots single out subject matters tailored to public interest.[119] In the second stage of gathering content, AI contributes through data mining. Algorithms are able to quickly and accurately process and analyze macro information to generate information, detect trends, add background context, and explore future prospects. In the current era of data overabundance, news outlets have shifted the focus from data collection to data representation, putting emphasis on selecting which data is worth pulling from.[120] AI's ability to not only filter data but also add value to specific points has served beneficial in this transition.[121]

AI has grown perilously important in the final stage of publication as well due to the rise of post-publication issues that clash with the media's ethical obligations. With an increase in use of social media and the internet as news sources, there has been an increasing demand for commentary moderation. Free access to online news platforms has provoked online harassment and spamming, hampering the media's healthy public discourse objective. But exhaustive moderation of thousands of comments takes a lot of resources to maintain.[122] AI takes this problem with the efficiency that no human can compare to in terms of both speed and volume. The Washington Post was the first news organization to integrate a commentary moderation AI system into its publication process. The Coral Project software "Talk with Modbot" successfully carried out this objective by constantly monitoring and filtering harmful comments on their online news stories. Ever since, other major media outlets such as The New York Times have followed, swiftly adapting AI systems for commentary moderation of their own.[123]

The production stage – the stage in which the news story is developed – requires journalists to find interesting angles, then formulate stories in a way that the public would consider both reliable and intriguing. The former requires objective storytelling skills as well as the incorporation of reliable and truthful sources, while the latter depends on creativity and

[119] Ibid., 6.
[120] Ibid., 4.
[121] Hada M. Sánchez Gonzales and Maria Sánchez González, "Bots as a News Service and Its Emotional Connection with Audiences. The Case of Politibot," Doxa Comunicación, Fall 2017.
[122] Miroshnichenko, "AI to Bypass Creativity", 7.
[123] Ibid., 8.

eloquence. Creativity in news production is essentially the ability to shape provided elements into a story in a way that executes the message best and that the audience will find entertaining to consume. This feature may seem secondary to serving journalistic ethical obligations; but the absence of creativity could lead to drawing less attention to stories with important news and information, interfering with the media's duty to inform the public. The necessity of creativity and its degree of significance in news stories will be further discussed when evaluating the qualitative competition between human and robot journalists.

Moreover, because the media runs under the business model, money is not out of the equation. If news outlets fail to sustain due to financial difficulties, the amount and variety of news provided would decline, hence stripping the public of much necessary information. This situation would not be ideal considering the public's dependence on the media as the fourth estate. But there are only so many ways that a news outlet could make money. First, there is membership, the only direct source of profit. Readers and viewers are customers, so the more they like the content that a news outlet produces, the more they would be willing to pay to access its stories or obtain special benefits limited to members. However, news outlets are increasingly abandoning the membership system to open up news and information to the entire public. Financial gain from membership thus contributes little to overall profit.

The majority of profit derives from indirect means such as ad revenue and hosting events. Evan Smith, the editor in chief of Texas Tribune, explained that hosting events at college campuses and beyond was "a cornerstone of Texas Tribune's financial success."[124] The ultimate grounds to gaining more eyes on ads and requests to host events is to increase viewership. News outlets can do so by not only constantly producing stories, but also ensuring that those stories are interesting enough to maintain the existing viewer base and attract new viewers. In doing so, they must still abide by their ethical obligation of public service, not simply choosing and producing stories that will get more attention but rather utilizing their journalistic judgement to determine what the public needs and how they should communicate it. Likewise, there is an emphasis on both the quantitative and qualitative aspect of news stories. AI participation looks different in the two areas.

Quantitative Competition

AI exceedingly wins the quantitative competition, which could be evaluated in terms of the ratio of time spent developing stories to the number of stories produced. The Washington Post's robot reporter published 850 articles in 2015, a feat that any human reporter would find hard to match.[125] Associated Press' Wordsmith increased the coverage of corporate earnings over tenfold by writing approximately 4,400 earnings stories per quarter. Despite the speed, these machines were able to carry out accurate information to the public. For robots, it takes seconds to pull statistics, compare them to previous data in similar contexts, incorporate the processed information in a news story template with smooth narrative structure, and publish the story.[126]

Genres of journalism including sports, business, finance, weather, and crime reporting are the first areas affected by AI's efficiency in speed. Thanks to the wealth of available statistics,

---

[124] Damaris Colhoun, "Three Ways News Outlets Are Making Money," Columbia Journalism Review, September 30, 2015, , https://www.cjr.org/analysis/the_new_models_new_models_for_news.php.
[125] Tatalovic, "AI Writing Bots," 2.
[126] Andrey Miroshnichenko, "AI to Bypass Creativity. Will Robots Replace Journalists? (The Answer Is "Yes")," Information 9, no. 7 (2018): 9, doi:10.3390/info9070183.

prescribed templates and styles, and advanced predictive models of these genres, computers easily replicate human journalists' reports.[127] For instance, The Guardian used a bot to combine game statistics and historical information of teams and players with pre-made phrases and connectors to compose stories. Associated Press collaborated with Automated Insights to develop a software that automated game recaps for Minor League Baseball, instantaneously covering all games of the season.[128]

Speed is relatively more emphasized in these genres, whether that be because of the quick updates sports fans expect or the of the need for communities to be informed about potential disasters or criminals nearby. The Los Angeles Times' algorithm Quakebot functions as a warning system reporting about California earthquakes.[129] Quakebot is able to quickly and accurately provide the public with knowledge about this natural disaster common in the state. Bot participation has also significantly changed traditional criminal coverage. It is impossible for human journalists to cover the myriad of crimes, so they must select a limited number of crimes they deem has the highest resonance potential. Contrarily, computers could process every single incident in the system and perform additional analysis such as categorizing race, gender, and neighborhood, creating secondary content of high importance that could have been overlooked. The Los Angeles Times' criminal reporting robot instantaneously processes data covering the vast city with a population of 10 million.[130] With such speed and scale in coverage, AI is able to serve the ethical obligation of public service more efficiently than any human. Because quantity is prioritized to quality in these genres where viewers look for numbers and facts rather than analysis, AI participation will only continue to grow in the upcoming years. Furthermore, some algorithms in these genres even proved successful in the qualitative realm; in 2012, AI system StatsMonkey wrote 1.5 million stories that viewers also found amusing to read, utilizing its unique trait of using baseball slang.[131]

In this manner, AI can produce news stories at a speed that human journalists can't compare to. So AI can produce more stories, but could it produce "better" stories that people find entertaining? More importantly, does it *need to* produce "better" stories in order to perform the ethical duties of journalism?

Qualitative Competition

In science and data journalism, the first and second stages of the journalistic process are most important since the stories focus heavily on data. The Science Surveyor project, developed at Columbia University, evaluates the trustworthiness of a research paper and then summarizes the key points.[132] These types of AI systems would automatically turn a research paper into a news story, possibly also providing context and background of the subject at hand while following the news-style template.[133] But human journalists can still provide benefits of their own in these genres. Freed from the tedious and time-consuming data processing stage, science

---

[127] Yair Galily, "Artificial Intelligence and Sports Journalism: Is It a Sweeping Change?" Technology in Society 54 (2018): 48, doi:10.1016/j.techsoc.2018.03.001.
[128] Ibid., 49.
[129] Hada M. Sánchez Gonzales and Maria Sánchez González. "Bots as a News Service", 66.
[130] Miroshnichenko, "AI to Bypass Creativity", 5.
[131] Miroshnichenko, "AI to Bypass Creativity", 9-10.
[132] Tatalovic, "AI Writing Bots," 4.
[133] Ibid., 5.

and data reporters can spend more time on performing in-depth investigations based on creative angles and in-person reporting on the field – both practices for which humans have an advantage, at least as of now.

Bots also contribute to quality information by reducing harmful content. AI develops stories at such a fast speed that they could develop multiple versions of a story, which editors could choose from. AI system SciNote was able to analyze science research papers and produce several versions of a news story that can be used to assemble a perfect publication for data.[134] The variety that AI provides decreases the possibility of publishing an inaccurate story. Factuality as well as fact-checking after production has become increasingly important due to the rise of fake news. Fake news – false information displayed under the facade of truthful journalism – is a great threat to the democratic media environment, as it goes directly against the journalistic ideal of pursuing truth to maintain trust between the media and the public. AI combats this issue in an efficient manner. AI is much less prone than humans to making mistakes and missing checks as long as it is initially programmed correctly and has no malfunction. They can also fact-check faster than humans; this is beneficial with long-form news stories that require lots of time to review, which human journalists may not be able to afford or prioritize.

AI promotes story quality in these manners, but can it produce stories that are just as appealing to read or view as those produced by human journalists, namely producing a news story creatively and artistically? Fields like investigative journalism as well as other more humanistic journalistic practices differ from the previously mentioned genres in that they are produced over time, including diverse sources and more thorough analysis.[135] Compared to others, these genres have to be more creative in its topic selection process – both a flexibility and a burden. In the assignment planning stage, humans have the unique advantage of serendipity. Characterized by "a strange combination of preparedness and impossibility to confidently prearrange," serendipity is impossible to calculate through some kind of linear process.[136] However, robots too have the ability to discover topics in a way that humans are incapable of. Humans tend to mistake correlation for causation due to their inability to analyze a great amount of data correlations by which you could at least confirm a higher statistical confidence. Algorithms, on the other hand, are able to calculate correlations amongst an enormous range of variables within seconds. Bots also don't try to search for causation in the first place, unlike humans that tend to draw incorrect conclusions about causation from imperfectly gathered data on correlations.[137] In addition, investigative journalism tends to be reactive rather than proactive due to its relatively complicated procedure. Machines could change this dynamic by generating leads, hunches, and anomalies based on data mining.[138] When successfully implemented, such algorithms could help bridge the knowledge gap in the proactive procedure.[139]

AI implementation has helped newsrooms meet their moral journalistic obligations in certain aspects. In the first two stages of journalistic process, AI implementation allows quicker,

---

[134] Tatalovic, "AI Writing Bots," 3.
[135] Meredith Broussard, "Artificial Intelligence for Investigative Reporting," Digital Journalism 3, no. 6 (November 28, 2014): 814-815, doi:10.1080/21670811.2014.985497.
[136] Miroshnichenko, "AI to Bypass Creativity", 14.
[137] Ibid., 15.
[138] Broussard, "Artificial Intelligence for Investigative Reporting", 815.
[139] Julian Dossett, "A Look Ahead: Where Artificial Intelligence May Take Journalism in 2019," Cision PR Newswire, November 20, 2018, https://mediablog.prnewswire.com/2018/11/20/where-artificial-intelligence-may-take-journalism-in-2019/.

broader, and more thorough data collection and analysis, which ultimately contributes to the goal of keeping the public well informed. Although quantity doesn't guarantee increased viewership, quantity does indeed increase the possibility of every citizen getting at least one story tailored to their interest. So how exactly do bots write or produce stories? Are they similar to those written by humans?

The Journalistic Turing Test

In 2015, NPR White House correspondent Scott Horsley performed a Journalistic Turing Test – a test comparing a machine's capacity against human intelligence – on the writing AI system Wordsmith.[140] Horsley and Wordsmith were to create a news story based on the earnings report of the Denny's restaurant chain. Horsley took seven minutes to complete the story whereas the robot took two minutes.[141] The results revealed key differences of news writing between the two. The robot's vocabulary was larger, as it compiles the entire dictionary. However, the robot produced a story that readers described as "dry" and "dull to read", having been programmed to use the most conventional words and sticking to the style most frequently used in similar news stories. Likewise, industrial specialization limits robots from using cooking or sports vocabulary in a financial report like human journalists would sometimes do as an artistic decision. Human writers have the freedom to use unconventional expressions that broaden context and make the story more entertaining. To produce an original style of writing, it is often necessary to deviate from rational necessity in this way; a good writer can use the "wrong" vocabulary and expressions deliberately.[142]

Correspondingly, when the two stories were presented to a group of editors without revealing which story was written by the robot or the human, they voted that the human's story excelled in categories of artfulness such as "well-written" and "pleasant to read," while the robot won categories necessary for hard news such as "objectivity", "clear description", and "accuracy." In this manner, human journalists exceeded in creativity, which makes a news story more engageable. Creativity is important in drawing an audience, improving the quality of a news story and hence, potentially increasing viewership even more than quantity. Under this assumption, human journalists may contribute more than AI to a newsroom's economic goal. But what's to say that AI can't eventually learn the syntax and phrases for a creative writing style most preferred by the public? NPR's Wordsmith was initially designed to mimic the straightforward tone of an AP news story. But after processing thousands of stories, it was able to mimic NPR's style and even sling its own breakfast-food metaphors.[143]

Furthermore, the categories in the Turing Test that the robot's story prospered in is more significant in achieving the ethical obligations of journalism. Factors like objectivity and accuracy are crucial for truthfulness, and a clear description will come across better to a public with varying levels of news literacy. At the end of the day, these categories will be prioritized over creativity in journalism because of the primary ethical objective and societal duty to truthful

---

[140] "The Turing Test, 1950," The Alan Turing Internet Scrapbook, https://www.turing.org.uk/scrapbook/test.html.
[141] Stacey Vanek Smith, "An NPR Reporter Raced A Machine To Write A News Story. Who Won?" NPR, May 20, 2015, https://www.npr.org/sections/money/2015/05/20/406484294/an-npr-reporter-raced-a-machine-to-write-a-news-story-who-won.
[142] Miroshnichenko, "AI to Bypass Creativity", 12.
[143] Smith, "An NPR Reporter Raced A Machine".

news reporting. Human journalists strive for truthfulness, but they are disposed to mistakes and personal agendas that cloud their decision. Contrarily, AI does not lie, has no personal attachment to any particular topics, or makes mistakes, as long as the data it is sourcing is truthful. These qualities will prosper over creativity in a democratic media environment that revolves around the public's trust in the media.

A potential concern is that a less creative story generated by a robot may be less appealing to an audience, leading to a decline in public engagement in news. This is unlikely, however, considering that not all human-produced news stories have thrived in creativity yet still gained attention because the public is ultimately interested in the story itself regardless of how the story is told. Thus, the goal of AI is to write a story that is not better than a human's but good enough to capture the essence of the news story. In fact, the editors that reviewed the two news stories in Horsley's Turing Test concluded that the distinction between the two was "so insignificant that they might as well have been written by different humans," and both stories "were acceptable for publication."[144] So maybe robots can't produce more creative stories than humans, but they can produce stories that the public will be just as receptive to. It is also possible that AI will learn creativity in the upcoming years. Even if bots can't recognize or produce beauty, they can calculate audience reactions to certain expressions, idioms, syntax structures, and learn to formulate aesthetically appealing stories accordingly.[145]

But the potential benefits that AI could provide have contradictory features that can actually corrupt the moral objective of journalism. Too many news stories may lead to overload, and the public will be forced to self-select what they need. Turning investigative journalism – a practice built on its long development procedure – into a proactive practice makes newsrooms vulnerable to faulty leads and hunches that could harm innocent people. Likewise, there are many ethical concerns regarding the implementation of artificial intelligence and algorithmic technology in journalism. In the following section, we outline ethical issues that are unique to the field of journalism, with reference to the media's moral duties in democratic societies.

Ethical Issues of AI Implementation in Journalism

Transparency and Accountability

The transparency ideal operates upon the promise that openness leads to security.[146] Although this ideal is simply an ideal and not necessary a reality, the ideal is highly preferred by participants of democratic societies where citizens politically participate based on their knowledge about what is happening in their societies. Because citizens obtain this knowledge primarily through the media, it is important for journalists to follow this ideal to maintain the public's trust. There are situations in which they will not. For instance, publishing information regarding war can be critical to national security. Another instance is when the exposure of a source could put the source in danger, even though the anonymity may make the news story less reliable. Plus, if journalists have to reveal their sources all the time, it would reduce the willingness of sources to provide information that may be of immense public interest. But

[144] Ibid., 14.
[145] Ibid., 15-16.
[146] Mike Ananny and Kate Crawford, "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability," New Media & Society 20, no. 3 (2016), doi:10.1177/1461444816676645.

excluding these exceptions, the public rightfully demands transparency from journalists. People would find news stories more reliable when they know where the information is coming from, how it was obtained, and how it was developed into a story.

The use of AI in the news gathering and producing process complicates procedures because of the limitations of algorithmic transparency: the disclosure of an algorithm design and functions. To achieve algorithmic transparency, AI technologists must explain to journalists the algorithm design as well as the logic behind the system. Such transparency is important to ensure that the technologists did not implement faulty ideas into the machine. Because AI can pick up inherent biases from the data they learn such as racism, sexism, or political favoritism, the use of a flawed AI would lead to the creation of a flawed news story.[147] Followingly, journalists using the software must fully comprehend the technological inner-workings of the algorithmic process themselves, in order to ensure they are using a safe program and then disclose the algorithmic procedure to the public.

But these AI systems are not objects of study that can be understood simply by scanning through the code.[148] To those not proficient in computer science, machine architecture, and computer codes – basically another language, algorithms will simply be gibberish. Thus, journalists are not likely to fully comprehend the technological side of the software they are using. This gives technologists, who are not bound by the ethical obligations of journalists, the power to manipulate the algorithm to serve their personal agenda, in which case would show through news stories produced through that algorithm. Furthermore, AI is a machine learning program; even an algorithm designed without bias and incorporating journalism's ethical obligations could evolve to develop flaws over time by pulling from faulty data and mimicking previous news stories that are prejudiced. Without knowledge about the algorithmic functioning, journalists would not be able to detect when this happens, which would lead to journalistic malpractice.[149] Journalists would have to constantly work with technologists to oversee AI functions by consistently deploying, configuring, and resisting problematic protocols.[150] This process would require time commitment and resources that many news outlets cannot afford.

Even if algorithmic transparency were achieved, we arrive at another question of who is to be the subject of blame when something goes wrong due to AI. The significance of transparency in journalism stems from the idea that being knowledgeable about a phenomenon creates opportunities to hold the responsible party accountable and request changes.[151] When citizens of democratic societies recognize a flaw in a news story, they hold journalists that were involved in the development of that story accountable. As a result, news outlets would punish those journalists through termination or suspension, and they would lose their reputation as reliable journalists. This system of accountability provides not only journalists the incentive to follow their journalistic ethical obligations to avoid these consequences, but also the public the sense of security from the fact that they have the power to hold journalists that violate those obligations accountable.

---

[147] Ovenden. "Will AI Destroy Journalism?"
[148] Mike Ananny and Kate Crawford. "Seeing without Knowing". 11.
[149] Mark Hansen et al., Artificial Intelligence: Practice and Implications for Journalism, Policy Exchange Forum, September 2017.
[150] Mike Ananny and Kate Crawford. "Seeing without Knowing". 11.
[151] Ibid., 2.

But when the flaw in the news story is caused by an algorithm, who should be held accountable? Technologists? Journalists? The entire news outlet? When multiple parties are involved, the power of accountability decreases; when accountability is distributed amongst all relevant parties, the degree of punishment is smaller for each party; and when accountability is concentrated on just one of the multiple parties, there is a sense of letting the other parties off the hook for something they should have been held accountable for. So there is no optimal situation to assign blame when multiple parties are involved. Overall, such breakdown of the accountability model could contribute in the decline of public trust in the media.

Data Protection and Privacy

The problem with accountability arises once again regarding concerns about AI invading privacy in the process of obtaining data. Under privacy and data protection laws, the news outlet as a whole would be held accountable as the publisher and data controller when an AI-produced news story is accused of having accessed data through unethical means.[152] The public would also find the fact that the direct source of the violations cannot be held accountable a reason to distrust that news outlet. As more media companies utilize AI and face these issues of accountability, the public would gradually grow skeptical of the media in general. Moreover, when an entire news outlet comes under fire for an AI's mistake, tanking their reputation as a journalism institution, all human journalists working for the company will also be negatively impacted with a defect on their own reputations.

Thankfully, content moderation by editors that review AI-produced stories before publication could mitigate privacy issues in the production and publication stage. If an AI-produced story contains private or confidential information about a claimant with such information not necessarily serving public interest, editors could simply take this portion out or stop the story from publication at all.[153] But human involvement in the data collection and analysis stage doesn't as easily resolve privacy concerns. An automated system is only as good as the data it accesses, so naturally we must make sure that the system is gathering data ethically and legally. But pinpointing whether AI inadequately processed personal data is difficult without comprehensive knowledge about the algorithmic function. As mentioned in the previous section, it is unlikely that human journalists will fully comprehend the technological side of AI. Hence, they will be prone to overlook the faulty ways through which the machine may have collected the data they use to develop news stories.

News Personalization

News personalization is news tailored to individuals based on their revealed interests. Personalization cannot be achieved without breaching an individual's privacy to a certain degree, since personalization is based on data gathered about the individual. Even when data is gathered within legal boundaries, it is ethically questionable to collect personal information about individuals without their consent. In 2018, it was revealed that Cambridge Analytica had

---

[152] "Robojournalism – Artificial Intelligence and the Media," Taylor Wessing, https://www.taylorwessing.com/download/article-robojournalism-ai-and-the-media.html.

[153] Stefan Hall, "Can You Tell If This Was Written by a Robot? 7 Challenges for AI in Journalism," World Economic Forum, January 15, 2018, https://www.weforum.org/agenda/2018/01/can-you-tell-if-this-article-was-written-by-a-robot-7-challenges-for-ai-in-journalism/.

amassed the personal data of millions of people through their Facebook profiles for political purposes. Although publishing personal information on social media was a decision made by the individuals, they were not aware that different companies would have access to and use their information. The incident resulted in Facebook's stock price dropping immensely and sparked a public outcry for tighter regulation of data usage.[154] This example demonstrates the way in which the public opposes such methods of data collection, regardless of the purpose. If news outlets were to use similar methods to personalize news, there is bound to be just as much backlash by the public, provoking skepticism towards the media.

But the greater issue of news personalization is its conflict with serving the public interest principle of ethical journalism. News personalization prevents exposure to diverse perspectives and instead provokes political polarization. "Purple", an app that uses bots to detect a readers' interest, individually delivers interest-aligning news stories through mobile phones.[155] Facebook also recently developed an algorithm that customizes news feeds according to individual users' reactions such as likes, shares, and posts. These machines have created an echo chamber in which people would grow stronger feelings about existing beliefs and remain ignorant about any other perspective. News outlets could maintain and even increase their viewership by automating tailored news articles for individuals, only presenting to them articles with ideas they agree with. However, by doing so, these media companies would essentially be feeding polarization, which directly goes against its ethical obligation to foster healthy public discourse.

Human Jobs at Stake

There are other ethical concerns that are not unique to journalism but very much apply in this industry as well. First, the introduction of AI can replace a lot of human journalists' jobs, leading to the decline in those journalists' well-beings. The last decade has already seen numbers of journalists dwindle dramatically across the globe. In the United States, they fell approximately one-third between 2006 and 2013; the United Kingdom and Australia have seen similar levels of decline.[156] AI will only contribute to this trend of decline for the journalistic profession.

Secondly, the inequality gap between smaller and larger newsrooms will widen with AI implementation. If AI can help news companies grow by increasing viewership, smaller news companies that don't have the money and resources to afford pricey AI developers would fall farther behind in the media market.[157] The exacerbation of asymmetrical power could lead to monopolies. The public would be resistant to such consequence, aware that the excessive power that few media companies have over public opinion gives those companies the power to manipulate news information in their favor. The economic disparity might only be temporary, as AI becomes more affordable over time. But in the following decade, the gap will inevitably widen with AI implementation. We can only hope that the public's trust in the media doesn't decline exponentially within that duration.

The Future of AI in Journalism

[154] Harry Davies, "Ted Cruz Campaign Using Firm That Harvested Data on Millions of Unwitting Facebook Users," The Guardian, December 11, 2015.
[155] Hada M. Sánchez Gonzales and Maria Sánchez González. "Bots as a News Service". 67.
[156] Ovenden. "Will AI Destroy Journalism?"
[157] Tatalovic, "AI Writing Bots," 2.

The media industry is already facing a crisis with declining public trust. In the United States, viewership fell for nearly every sector of news media in 2017.[158] Those who unsubscribed to their previous news sources commented that their increased skepticism paralleled to the increase of fake news on online platforms.[159] Different democratic societies had varying reasons for decreased viewership, but they all related to wariness about the new era of technology. Citizens have grown fearful from an inability to distinguish reliable and unreliable news sources. The unregulated nature of new technology has fed this fear. The provision of poor quality information from a failure to manage AI will participate to this pattern.[160]

AI has been progressively utilized by media outlets in the past decade, mainly those run by big-money companies that have the resources to do so. Considering the economic and performative benefits that AI presents, the numbers will only continue to rise in the following decade. In fact, AI usage will likely increase dramatically as the technology becomes more accessible. Yet, few news organizations are currently thinking of a game plan for journalism in the coming age of automation. According to Amy Webb's 2017 Global Survey of Journalism's Futures, nearly 70 percent of the surveyed journalists claimed that their newsrooms are not conducting analysis of emerging AI trends and how AI will impact their newsrooms in the next 5 to 10 years.[161] This must change. Considering the exponential growth of AI in the industry, there is an urgent need for journalists to prepare themselves for AI incorporation. Newsrooms – both those currently utilizing AI and those not – should brainstorm detailed methods through which AI could participate without moral conflict. With that, we provide some recommendations for the ethical usage of AI in journalism going forward.

Ethics Boards and Guidelines

In order to ensure that newsrooms are incorporating AI ethically, they must establish an ethics board that keeps them in check. Following the acquisition of AI system DeepMind, Google established an ethics board in collaboration with the developers of DeepMind and other AI researchers.[162] Since Google's definition of ethics may not parallel the rest of the world's, the company gave their shareholders as well as DeepMind the authority to prompt legal action against them when they violate the terms provided through the ethics board.[163] In this manner, every media company's ethics board must have the power to monitor and regulate AI utilization in the newsroom. Without such authority, the board could easily become just a public relations ploy.

The Google ethics board created five policies detailing how to ethically go about AI application. The policies provided guidelines to determine factors such as when it would be "ethical for systems to cause physical harm to humans," how to limit "the psychological

---

[158] Michael Barthel, "5 Facts about the State of the News Media in 2017," Pew Research Center, August 21, 2018, , https://www.pewresearch.org/fact-tank/2018/08/21/5-facts-about-the-state-of-the-news-media-in-2017/.
[159] "Freedom of The Press," Freedom House, 2017, https://freedomhouse.org/report/freedom-press/freedom-press-2017.
[160] Paul Chadwick, "As Technology Develops, so Must Journalists' Codes of Ethics," The Guardian, January 21, 2018, , https://www.theguardian.com/commentisfree/2018/jan/21/technology-codes-ethics-ai-artificial-intelligence.
[161] Amy Webb, Global Survey of Journalism's Futures, issue brief no. 106, Future Today Institute (2017).
[162] "DeepMind Ethics & Society Principles," DeepMind, https://deepmind.com/applied/deepmind-ethics-society/principles/.
[163] Ovenden. "Will AI Destroy Journalism?"

manipulation of humans" and how to prevent "the concentration of excessive power."[164] In this light, we propose policies of our own – an ethical guideline for newsrooms to utilize when incorporating AI systems. Based on the ethical analysis we've performed throughout this chapter, we suggest that every news outlet must:

- Develop a formal educational procedure with the help of the AI developers. Every human journalist in the newsroom must go through the procedure to learn about the AI system they will be using. The procedure must be detailed, structured, and comprehensive. Journalists should be tested on the content after going through the procedure to ensure that they have truly comprehended the functions of the machine.
- Publicly disclose which parts of their news developing process AI is involved in.
- Publicly disclose the AI's algorithmic procedure, attempting to achieve algorithmic transparency.
- Develop a clear system of accountability in which all journalists and technologists will be held liable in the case that AI causes ethical problems. The punishment for overlooking mistakes caused by AI or programming faulty ideals to AI systems must be great enough to provide every involved party the incentive to create or use an algorithm that is ethical.
- Not fire a human journalist on the grounds that the AI is able to perform all the responsibilities that they are. It is impossible that the human journalist has absolutely *nothing* to offer, considering that they still possess uniquely human advantages like serendipity and creativity.
- Not use AI for personalizing news.
- Not use AI to gather data including personal information about individuals without their consent. The only exception would be when such information belongs to a public figure and the information holds critical value that will serve public interest.
- Always copy-edit and review an AI-produced news story before publication. If the news story contains any ethically questionable content, the editor that was responsible for reviewing that news story will be held most accountable.
- Constantly monitor the AI's data collection process as well as function, ensuring that the data was harvested through ethical means and that the machine has not evolved to cause ethically concerning problems.
- Train lower-level journalists on the use of AI as well as the traditional ethical duties of the media, preparing them to become editors themselves. Through this training procedure, lower-level journalists must prepare themselves to copy-edit and review news stories involving AI assistance or produced solely by AI.

Human Involvement

Even if these ethics boards and guidelines are established to manage AI, human involvement will continue to be necessary. The greatest danger of AI is arguably its unpredictability. Unlike traditional machines, AI has the ability to evolve. We do not have enough empirical evidence on AI's functioning in journalism to blindly trust it to operate without human oversight. So for the sake of safety, human journalists should continue to serve their original roles in newsrooms to the extent they currently do, at least for the following decade. Beyond 10 years, lower-level journalists may become less useful to newsrooms as AI serves their

[164] Ibid.

functions. But editors must always remain in place. Because AI behavior can neither be easily predicted nor trusted to be held constant, human journalists must keep AI in checks with risks of unpredictability in mind.[165] AI should essentially be treated like entry-level journalists that need guidance and checkups. No matter how good of a story an entry-level journalist produces, their work will go through a copy-editing process afterwards. In the same way, human involvement is crucial in every step of the journalistic procedure regardless of AI's competency.

In the assignment planning stage, assignment editors should review how the AI system selected the topic and evaluate whether the topic is apt for the current climate. In the content gathering stage, editors must review whether the machine gathered data ethically, without making any privacy violations, and make sure that the sources it drew data from are reliable. Reiteratively, this requires human journalists to be fully knowledgeable about the AI system's functioning. In the production stage, editors should copyedit stories produced by AI, fact-checking and filtering any wrongful ethical implications in the story, just as they would with any other human-produced news story. Additionally, editors should publicly reveal how the AI functions and in which stages of the journalistic procedure AI was involved, for the sake of transparency and maintaining public trust.

This doesn't mean that after 10 years, we should only hire editors. There are certain features like serendipity and creativity that humans excel at. A serendipitously conceived topic as well as an artistically formulated story can arise from an amateur journalist just as much as it could from a senior editor. It's true that amateur journalists' positions will be the first at stake when AI becomes more prevalently utilized and able to perform the tasks they do, if not, better. But we must recognize that all editors start as amateurs too. If anything, having started in the newsroom when AI usage is already commonplace, future entry-level journalists will learn about AI incorporation in the journalistic process from scratch. When these amateurs become editors, they will be even more proficient in navigating the ethical use of robots in their newsrooms than existing editors today. From an economic standpoint, newsrooms may find this kind of human involvement excessive and a waste of resources. Nevertheless, from an ethical standpoint, human involvement is fundamental in achieving the moral duties of journalism, which should be prioritized considering the momentous role that the media plays in democratic societies.

[165] Mike Ananny and Kate Crawford. "Seeing without Knowing". 13.

Works Cited

Ananny, Mike, and Kate Crawford. "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability." *New Media & Society* 20, no. 3 (2016): 973-89. doi:10.1177/1461444816676645.

Barthel, Michael. "5 Facts about the State of the News Media in 2017." *Pew Research Center*, August 21, 2018. https://www.pewresearch.org/fact-tank/2018/08/21/5-facts-about-the-state-of-the-news-media-in-2017/.

Berry, Stephen J. "Why Objectivity Still Matters." *Nieman Reports* 59, no. 2 (Summer 2015). https://niemanreports.org/articles/why-objectivity-still-matters/.

Broussard, Meredith. "Artificial Intelligence for Investigative Reporting." *Digital Journalism* 3, no. 6 (November 28, 2014): 814-31. doi:10.1080/21670811.2014.985497.

Chadwick, Paul. "As Technology Develops, so Must Journalists' Codes of Ethics." *The Guardian*, January 21, 2018. https://www.theguardian.com/commentisfree/2018/jan/21/technology-codes-ethics-ai-artificial-intelligence.

Colhoun, Damaris. "Three Ways News Outlets Are Making Money." *Columbia Journalism Review*, September 30, 2015. https://www.cjr.org/analysis/the_new_models_new_models_for_news.php.

Cunningham, Brent. "Re-thinking Objectivity." *Columbia Journalism Review* 42, no. 2 (July/August 2003): 24. https://archives.cjr.org/feature/rethinking_objectivity.php.

Davies, Harry. "Ted Cruz Campaign Using Firm That Harvested Data on Millions of Unwitting Facebook Users." *The Guardian*, December 11, 2015.

"DeepMind Ethics & Society Principles." DeepMind. https://deepmind.com/applied/deepmind-ethics-society/principles/.

Dossett, Julian. "A Look Ahead: Where Artificial Intelligence May Take Journalism in 2019." *Cision PR Newswire*, November 20, 2018. https://mediablog.prnewswire.com/2018/11/20/where-artificial-intelligence-may-take-journalism-in-2019/.

Fox, Carl. "Public Reason, Objectivity, and Journalism in Liberal Democratic Societies." *Res Publica* 19, no. 3 (2013): 257-73. doi:10.1007/s11158-013-9226-6.

"Freedom of The Press." Freedom House. 2017. https://freedomhouse.org/report/freedom-press/freedom-press-2017.

Galily, Yair. "Artificial Intelligence and Sports Journalism: Is It a Sweeping Change?" *Technology in Society* 54 (2018): 47-51. doi:10.1016/j.techsoc.2018.03.001.

Gonzales, Hada M. Sánchez, and Maria Sánchez González. "Bots as a News Service and Its Emotional Connection with Audiences. The Case of Politibot." *Doxa Comunicación*, Fall 2017, 64-67.

Hall, Stefan. "Can You Tell If This Was Written by a Robot? 7 Challenges for AI in Journalism." World Economic Forum. January 15, 2018. https://www.weforum.org/agenda/2018/01/can-you-tell-if-this-article-was-written-by-a-robot-7-challenges-for-ai-in-journalism/.

Hansen, Mark, Meritxell Roca-Sales, Jon Keegan, and George King. *Artificial Intelligence: Practice and Implications for Journalism*. Policy Exchange Forum. September 2017.

Jacquette, Dale. *JOURNALISTIC ETHICS: Moral Responsibility in the Media*. ROUTLEDGE, 2017.

Jansen, Sue Curry. *Censorship: The Knot That Binds Power and Knowledge*. New York: Oxford University Press, 2010.

Milton, John. *Areopagitica, 1644*. New York: Published by Columbia University Press for the Facsimile Text Society, 1927.

Miroshnichenko, Andrey. "AI to Bypass Creativity. Will Robots Replace Journalists? (The Answer Is "Yes")." *Information* 9, no. 7 (2018): 183. doi:10.3390/info9070183.

Ovenden, James. "Will AI Destroy Journalism? The Profession Is Struggling and AI Could Be the Final Nail in the Coffin." Innovation Enterprise, December 21, 2017. https://channels.theinnovationenterprise.com/articles/will-ai-destroy-journalists.

"Robojournalism – Artificial Intelligence and the Media." Taylor Wessing. https://www.taylorwessing.com/download/article-robojournalism-ai-and-the-media.html.

Schauer, Frederick. "Facts and the First Amendment." *UCLA Law Review* 57, no. 4 (2014): 901.

Smith, Adam. *The Wealth of Nations*. New York: American Home Library, 1902.

Smith, Stacey Vanek. "An NPR Reporter Raced A Machine To Write A News Story. Who Won?" *NPR*, May 20, 2015. https://www.npr.org/sections/money/2015/05/20/406484294/an-npr-reporter-raced-a-machine-to-write-a-news-story-who-won.

Tatalovic, Mico. "AI Writing Bots Are about to Revolutionise Science Journalism: We Must Shape How This Is Done." *Journal of Science Communication* 17, no. 01 (2018): 2. doi:https://doi.org/10.22323/2.17010501.

"The Turing Test, 1950." The Alan Turing Internet Scrapbook. https://www.turing.org.uk/scrapbook/test.html

Webb, Amy. *Global Survey of Journalism's Futures*. Issue brief no. 106. Future Today Institute. 2017.

Ethical Concerns of Targeted Advertising

Jillian Kohn

In 2012, news of a father berating a Target employee for sending his daughter coupons for diapers due to his belief that the company assumed she was pregnant, was met with shock, unease, and concerns over privacy. The father eventually apologized to Target upon learning that his sixteen-year-old daughter was, in fact, expecting. In the *New York Times* article *How Companies Learn Your Secrets* published in the same year, Andrew Pole a statistician previously employed by Target relates to author Charles Duhigg just how Target was able to recommend and suggest products to its customers through data on past purchases. For expecting mothers, Pole could assign a "pregnancy prediction score" and even estimate a woman's due date based off the time and frequency with which she purchased items such as soap and cotton balls.

Throughout time, customer data has been used to predict future purchases and, in some cases, even shape consumer choice. In the last decade, artificial intelligence (AI) has vastly improved the predictive capacity of consumer data, largely due to AI's potential to sort and categorize large amounts of data with greater accuracy and speed. While exciting in its potential, the growing use of AI in marketing and advertising raises concerns about consumer privacy. The scale and accuracy with which artificial intelligence can forecast consumer choices, due to the vast amount of data and speed with which it sorts and categorizes that data to craft customer segments, raises concerns about where these practices could lead, "but with their analysis moving into areas as sensitive as pregnancy, and so accurately, who knows how else they might start profiling Target shoppers?"[166] The main concern involves the ability for Target to show they are aware of certain private attributes that customers have not disclosed to the company or perhaps even to close family and friends. Regarding this concern, Pole says "if we send someone a catalog and say, 'Congratulations on your first child!' and they've never told us they're pregnant, that's going to make some people uncomfortable...We are very conservative about compliance with all privacy laws. But even if you're following the law, you can do things where people get queasy."[167] This interaction points to the ethical and legal concerns in the use of artificial intelligence in targeted advertising. It is not necessarily the artificial intelligence itself that raises these concerns, but how it is programmed and employed by individuals and companies.

Although it might be legal, the use of artificial intelligence in advertising raises a key ethical concern of the fragile, unpredictable, and unknown nature of data privacy. Target executives are primarily concerned with maintaining the fine line between protecting data and exploiting it, in order to continue to both maximize profits and to avoid a public relations disaster. By claiming that they are in compliance with all privacy laws and regulations, they are implying that they are protecting all sensitive information such as health information. However, it is not only the stated policies but the level to which these policies are upheld and enforced that more accurately reflects how well data is protected.

---

[166] Hill, Kashmir. "How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did." Forbes. https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#1d3ac9136668

[167] Ibid.

Consumer choice has always been based on needs (and, as income grows, wants) of the customers. Marketing interacts with customer choice by creating value for the customer and in turn, capturing value from customers in the form of sales and profits. To market products effectively, a company must understand the customers' needs in order to provide them with products of superior value, and then distribute and promote these products well. Customer choice is the end result, when customers choose to engage with the company by purchasing a product. The desire to understand the customer's needs and then provide customers with products of superior value and then distribute and promote them well, leads to the customers choosing to engage with the company in the form of purchases.

The digital age has transformed the way that companies operate their marketing strategy in order to engage consumers. The internet allows for direct communication, market research, and distribution to form a relationship between the two. Instead of identifying and then searching for potential customers, advertisers can know intimate details about potential customers as well as the likelihood that they will become customers. The use of personalized content recommendation algorithms or simply targeted ads are used to better match the needs of customers to what the company can provide. This algorithmic form of market research raises multiple ethical concerns about the use of artificial intelligence in marketing and its effects on data privacy, equality, freedom, and consumer choice. The ease with which the processes gather, sort, and analyze data creates an efficient, cost-effective, and better matched ad for consumers, yet not all consumers are keen to the idea of their personal data, posts, and decisions, being shared across multiple companies and platforms.

Many of these concerns mirror those of artificial intelligence in other industries, but there is less available knowledge on how artificial intelligence has disrupted these industries due to their novelty. Artificial intelligence in advertising, however, has had to adapt to complaints and navigate the ethical challenges it has encountered. Therefore, other industries that are considering adopting the use of similar technology can look to advertising policies' strengths, weaknesses, and opportunities. When looking to utilize machine learning in any industry there is a need to be proactive about mitigating potential ethical violations.

Companies such as Facebook and Google have been employing targeted ads for years, but more recently companies such as Amazon and IBM have too adopted new uses for machine learning in their advertising decisions. These include Amazon's free product samples as well as IBM's Watson for Advertising (formerly known as The Weather Company). The use of artificial intelligence to find the right customers is widespread. According to Mr. Duhigg, "almost every major retailer, from grocery chains to investment banks to the U.S. Postal Service, has a 'predictive analytics' department devoted to understanding not just consumers' shopping habits but also their personal habits, so as to more efficiently market to them."[168] Companies utilize artificial intelligence to more accurately and efficiently identify who will most likely become or remain a customer or who has the potential to become a customer and allows those customers to connect with businesses in a more meaningful way, "for companies like Target, the exhaustive rendering of our conscious and unconscious patterns into data sets and algorithms has revolutionized what they know about us and, therefore, how precisely they can sell."[169]

---

[168] Duhigg, Charles. "How Companies Learn Your Secrets." NY Times. February, 16, 2012. https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=1&hp

[169] Ibid.

Google and Facebook, the two largest companies offering targeted advertising to their business partners, provide multiple ways to show users ads. Advertisers tell Facebook what their target audience is and Facebook in turn matches the ad to those it believes will be interested in the ad based off of users' activity. Targeted ads utilize the data of users' activity across Facebook and products such as "pages you and your friends like, information from your Facebook and Instagram profile, and places you check in using Facebook."[170] Similar to how Facebook utilizes data to craft customer profiles is Target's Guest ID. Linked to the Guest ID is demographic information such as "your age, whether you are married and have kids, which part of town you live in, how long it takes you to drive to the store, your estimated salary, whether you've moved recently, what credit cards you carry in your wallet and what websites you visit."[171] All of this data is discerned from information that users input themselves and from their purchasing history. Additionally, Target can buy data about your "ethnicity, job history, the magazines you read, if you've ever declared bankruptcy or got divorced, the year you bought (or lost) your house, where you went to college, what kinds of topics you talk about online, whether you prefer certain brands of coffee, paper towels, cereal or applesauce, your political leanings, reading habits, charitable giving and the number of cars you own".[172] This data is incredibly specific and Target has the potential to use it to create personalized ads for customers, and yet the retailer has declined to comment on what demographic information it collects or purchases. This lack of transparency only fuels the concerns over use of personal information, where it is being used, when, and by whom.

An additional process by which Facebook acquires user data is through users' activity with other businesses that then upload customer lists with phone numbers and emails that can be matched to Facebook profiles. This activity can involve signing up for newsletters or coupons and discounts at a business or simply making a purchase. More targeted methods include gathering data on users' activity on other websites and apps. These sites utilize pixels that track when a user is searching for, purchasing, or viewing a certain product, researching, or downloading a certain company's app. Through Facebook Pixel, these websites can send Facebook user information directly to then match ads to products or services the user has previously searched for.

Location is also a critical part to how these companies utilize a user's data. When a user connects to the internet, uses Facebook or Instagram, or is simply on their phone, they are tracked to provide targeted ads to match to the users' whereabouts. This location tracking provides businesses a way to target those that are around them or going to be at a given time. Facebook claims that protecting users' privacy is central to how they've designed their ad system and one of the key ways they've claimed to be able to do so is by allowing users "to have a say in the ads they see."[173] Having privacy as a main concern has forced Facebook to create an ad system in which advertisers can glean and utilize all the aforementioned user data without learning who individuals are. Essentially, Facebook claims that by protecting specific sensitive information like a users' name or selling user information that could identify users, they are fulfilling their requirement to protect their users' data.

---

[170] About Facebook Ads. https://www.facebook.com/ads/about.

[171] Hill, Kashmir. "How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did."

[172] Ibid.

[173] About Facebook Ads. https://www.facebook.com/ads/about.

Google identifies information that should be handled with care including full name, email address, mailing address, phone number, national identity, pension, social security, tax ID, health care, or driver's license number, birth date or mother's maiden name, financial status, political affiliation, sexual orientation, race or ethnicity, and religion.[174] This managing of ad preferences comes in the form of specific controls users have to help determine what ads they see such as hiding an ad, hiding all ads from a specific advertiser, inquiring for more information about the appearance of the ad, reporting a problem with an ad, and reviewing and changing the things that influence what ads users see in their ad preferences. The previously mentioned example involving retailers such as Target highlight potential problems with privacy. The privacy violation results from a spillover of the provided personal information similar to the information that identifies as having to be handled with care.

Privacy

Privacy is typically thought of as the right to be left alone, but when it comes to the use of online and digital advertising, privacy is more relevant to protecting and choosing to withhold personal information. Privacy as the right to selectively express oneself and choose to seclude personal information about oneself, is a key concern of the use of artificial intelligence in advertising. As outlined in the aforementioned Target example, the main ethical concerns raised by the use of targeted advertising involve privacy. What information can be accessed and utilized by artificial intelligence and how it will be used and who will have access remain the key privacy considerations. People may be unaware that they are actively providing companies and advertising agencies with information that can be traced back to them when they sign up for a social media site or service. Furthermore, companies and the artificial intelligence that they use to gather and gain insights from data can be hacked and misused, causing users to question the security of the technology and whether their information is actually being protected to the best of the company's ability. It is not the fact of users doubting security that is the problem, but rather the potential security issues.

IBM Watson uses a client-first approach to address these issues of privacy. It involves having clients build their own artificial intelligence to ensure that data remains proprietary and secure.[175] With this principle acting as their foundation, the main benefits can include results that are "trusted, insight-driven solutions, that brands can leverage across their entire marketing strategy to help save time, money and build more meaningful customer connections".[176] Simply from statements such as these available on their website, IBM has proven to be far more proactive and transparent than Facebook and Amazon when it comes to how data is being used, what their ad policies stated goals are, and how they plan to mitigate risks of breaches in data privacy.

The main concerns of data gathered and insights constructed through artificial intelligence are who eventually gets access to the data, for how long, and what they will do with it. The user may not know the answer to any of these questions because many of these may not be known to the user, they did not consent prior to its use, or they did not know what they were

[174] Google Advertising Policies

[175] AI for Advertising. IBM. https://www.ibm.com/watson-advertising

[176] Ibid.

consenting to. Users need to be aware of where their data is actually being used and so they can make an informed decision on whether they want to supply certain personal information. This can be difficult to ascertain with the use of artificial intelligence because the "use of existing data sets may be reaching a point where data can be used and recombined in ways that people creating that data in, say, 2000 or 2005, could not reasonably have foreseen or incorporated into their decision making at the time."[177] People provide a wide range of personal information to carry out rather mundane tasks and are not necessarily aware of how it will be used by third-parties, in the future, or both. The ethical and legal concerns are for this massive aggregation of data points such as zip code or age that in the past individually would not have been able to identify users can easily be done with the efficiency of today's algorithms.[178] The unknown nature of companies' use of these algorithms may change how users' view their privacy settings.

One such use that may cause users to want to change their privacy settings for their ads is when artificial intelligence becomes able to identify individuals, thus potentially breaching assumptions of anonymity. The uncertainty of the future of their data and the firms that will have access to their data and the potential uses for it, can cause users to become unsure of their decisions of sharing their data. They may feel as if they do not have a choice in the matter since very mundane tasks such as subscribing to an online newsletter or signing up for a social networking site requires the relinquishment of personal and contact information. Questions remain on the overall ethics of retaining users' data. Potential regulations could restrict the length of time that companies can retain certain data, but whether that would be beneficial or detrimental long-term is case-specific and unclear in the present. Present concerns revolve around privacy; therefore, regulations on the restriction for data reuse may be necessary. It may also be helpful to identify what areas of data should explicitly not be allowed to be reused either because the effects are unknown, unpredictable, or violate other ethical considerations.

Additionally, when an algorithm allows for data to be used to discriminate against a user, especially if they did not explicitly consent to its collection in the first place, it becomes a privacy issue. This data-based discrimination from algorithmic outcomes occurs "because the algorithm itself will learn to be biased on the basis of the behavioral data that feeds it."[179] For example, a study in 2015 found that users that Google had identified as female, received fewer ads for an executive coaching service.[180] Google ads showed an advertisement for a career coaching service that promised larger salaries more frequently to male-identified users than female-identified users.[181] Whether this means that the algorithm has learned to be biased and therefore only shows a certain audience an ad is unclear. However, the use of a person's data to discriminate against them raises concerns of equality of artificial intelligence in advertising.

Equality

[177] Tucker, Catherine. "Privacy, Algorithms, and Artificial Intelligence." pg. 12

[178] Ibid.

[179] O'Neil, C. (2017). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy." Broadway Books.

[180] Tucker, Catherine. "Privacy, Algorithms, and Artificial Intelligence." pg. 15.

[181] Datta, A., M. C. Tschantz, and A. Datta (2015). "Automated Experiments on Ad Privacy Settings. Proceedings on Privacy Enhancing Technologies". 2015(1), 92–112. https://www.nber.org/chapters/c14011.pdf.

One of the main and perhaps most salient ethical concerns of targeted advertising is discrimination, "an action can be found discriminatory, for example, solely from its effect on a protected class of people, even if made on the basis of conclusive, scrutable and well-founded evidence."[182] The concern with targeted advertising is that it lends itself to implicit discrimination. Although anti-discrimination laws prevent explicit statements that advertise for certain candidates on the basis of their race, religion, or country of origin, and then amended to include gender, disability, and familial status targeted ads have the ability to be shown only to groups of people with certain demographic characteristics. Prior to artificial intelligence, people had more choice over the content the consumed, albeit not full choice. Advertisers still intervened and could choose where to put their ads, such as in particular TV channels, neighborhoods, cities, regions, and magazines, on the basis of consumer demographics, it was just less individually targeted than with the introduction of artificial intelligence. Now with its advanced targeting, Facebook, gets to make those content choices for users. Marketers need to keep in mind the ethical implications of what showing one group certain content while excluding another means. These groups will not only not have access to these ads, but it allows those publishing the ads for housing or employment to exclude these groups from these housing or employment options. This is keeping people out of an ad, rather than targeting to get them in front of the ad. Keeping someone out of an ad is of use to advertisers if they would like to prevent them from viewing it. These advertisers don't want certain groups to see their ads because they don't want the ad to work on them. This is the opposite of concerns of people viewing certain ads, the main purpose of targeted advertising. The ability to advertise in this way has increased due to the mechanism of targeted ads that identify specific limited audiences based on either the same or similar demographic and lifestyle factors that are used to discriminate.

This violation of equality occurs when artificial intelligence is used to exclude certain groups of people while benefiting others. This could result in an even larger gap in discrimination between people based on factors such as income, race, or location. Facebook has dealt with a case involving advertisers being able to discriminate against African American, Asian American, and Hispanic populations in real estate. ProPublica was able to buy ads for real estate that blocked users with an "affinity" for these populations. They purchased ads targeted at users that were identified as house hunting by Facebook and excluded anyone with an "affinity" for African American, Asian-American, and Hispanic people. As confirmed by civil rights attorney John Relman, this is a blatant violation of the Fair Housing Act of 1968 which makes it "illegal to make, print, or publish, or cause to be made, printed, or published any notice, statement, or advertisement, with respect to the sale or rental of a dwelling that indicates any preference, limitation, or discrimination based on race, color, religion, sex, handicap, familial status, or national origin"[183] with violations resulting in potential fines of thousands of dollars. The actual discrimination concern was that the ad platform had the ability to exclude people

[182] Mittelstadt, Brent Daniel, et al. "The Ethics of Algorithms: Mapping the Debate." *Big Data & Society*, Dec. 2016, doi:10.1177/2053951716679679. pg. 5.

[183] Parris, Terry and Angwin, Julia. "Facebook Lets Advertisers Exclude Users by Race." ProPublica. https: //www.propublica.org/article/facebooklets-advertisers-exclude-users-by-race, 2016.

based on their "ethnic affinity" (a curated attribute) when targeting ads related to housing.[184] Facebook does not ask its members about race because ads that exclude people based on race, gender and other sensitive factors are prohibited by federal law in housing and employment. Instead, Facebook assigns members an "ethnic affinity" based on pages and posts they have liked or engaged with on Facebook to place them into specific groups that advertisers can then exclude.[185] Facebook responded by banning the use of ethnic affinity attribute for certain types of ads. Steve Satterfield, the director of privacy and public policy at Facebook, offered a way in which excluding certain groups of users on the basis of race would be beneficial for an advertiser, "might run one campaign in English that excludes the Hispanic affinity group to see how well the campaign performs against running that ad campaign in Spanish. This is a common practice in the industry."[186] This may be true; however, it has not prevented advertisers from utilizing Facebook's advertising tools to prevent some users based on their assumed race from viewing certain ads.

On March 19, 2019, five discrimination lawsuits filed from November 2016 to September 2018 between civil rights advocates and Facebook were settled, resulting in changes that will prevent "advertisers for housing, employment or credit from discriminating based on race, national origin, ethnicity, age, sex, sexual orientation, disability, family status, or other characteristics covered by federal, state, and local civil rights laws"[187] on Facebook, Instagram, and Messenger.[188] The lawsuits included ones filed by the National Fair Housing Alliance, Communications Workers of America, individual consumers and job seekers represented by the ACLU and numerous other law firms, and several regional fair housing organizations. Facebook has published a new ads policy in response to multiple ProPublica articles that exposed the discrimination against the people who are not targeted for real estate. The recent updated ads policy includes over 1,000 more employees to review ads and prohibits ads promoting categories such as cryptocurrency that are typically associated with misleading and deceptive advertising practices.[189]

The settlement resulted in changes including the establishment of a new separate advertising portal which will have limited targeting options for creating housing, employment, and credit ads. Key rules will include the removal of targeting by zip code as well as of gender, age, and multicultural affinity targeting options, a minimum geographic radius of fifteen miles from a specific address or city center. Secondly, educational and training materials will be developed in tandem with the Plaintiffs and they will be allowed to test the success of Facebook's implemented terms of the settlement.

---

[184] Friedler, Sorelle A. and Wilson, Christo. "Potential for Discrimination in Online Targeted Advertising." Proceedings of Machine Learning Research 81: 1-15, 2018.
http://proceedings.mlr.press/v81/speicher18a/speicher18a.pdf

[185] Parris, Terry and Angwin, Julia. 2016.

[186] Ibid.

[187] *Summary of Settlements Between Civil Rights Advocates and Facebook.* ACLU. March 29, 2019.
https://www.aclu.org/other/summary-settlements-between-civil-rights-advocates-and-facebook

[188] Ibid.

[189] Leathern, Rob. "New Ads Policy: Improving Integrity and Security of Financial Product and Services Ads." Facebook Business. January 30, 2018.
https://www.facebook.com/business/news/new-ads-policy-improving-integrity-and-security-of-financial-product-and-services-ads.

Interestingly, Facebook is also employing artificial intelligence in order to catch potentially discriminatory ads and has stated that they are working to improve that machine learning's ability to review ads. On the human side, Facebook is educating advertisers about what is potentially discriminatory or in violation of the community standards by providing a prompt with a reminder of the company's anti-discriminatory policies.[190] Lastly, Facebook requires advertisers that they identify as having ads that offer housing, employment or credit, to certify that they are complying with their anti-discrimination policy, and claim that they already reject thousands of these types of ads a day.[191] These measures have been put into place to prevent ads that violate Facebook's policies before they run however still may not entirely eradicate this issue since advertisers must self-police initially. The line between targeting users and excluding users through these advertisements can at times be difficult to discern. However, when the algorithm is utilized unethically to exclude groups of people from exposure to certain ads, ethical violations occur. Similarly, an additional ethical concern of artificial intelligence in targeted advertising results from the users' actual ability to choose what ads they see and what purchases can be made from them.

Freedom

Freedom is the unrestrained right to act as one so chooses. In order to do this, choices must be readily available and the liberty to choose between them unhindered. The choice of what ads users see and whether they have the ability to protect their own privacy, however, requires transparency of what data is being collected, and how that data leads to the ads they see, and if they can decide how that data is being used. This choice allows them to control what sensitive information they want to be protected. The practice should not have ethical violations so long as companies continue to allow customers to opt in or out of advertising and modify their ad preferences. This means the privacy concerns that arise due to worry about sites gaining personal information can be mitigated and those that are annoyed by the ads do not have to see them. If there is an effect on the ads that they see when users update their interests and demographic information, then they will have a level of ad choice. This can be difficult to measure. If a user removes an interest that is being used to inferred someone's ad preferences, then in order to be in compliance with the stated preference policy such as Google's Ad Settings, the number of ads shown related to that interest should decrease.

Recently, a more explicit yet similar service to Target's coupons has been rolled out by Amazon in the form of free samples. This is an automatic service and Amazon customers do not opt-in to it upfront. The online retailer is now allowing brands to pay to have their products shipped to customers that Amazon chooses to receive them. Amazon decides who to send the products to based off previous purchases and what its algorithm believes people will like and therefore buy as a result. Brands can pay Amazon so that their products can reach certain customers. Amazon's website says of the pilot program, "Amazon surprises select customers with samples that we think will be delightful and helpful…Amazon helps you discover products

[190] *Reviewing Targeting to Ensure Advertising Is Safe and Civil.* Facebook Business. https://www.facebook.com/business/news/reviewing-targeting-to-ensure-advertising-is-safe-and-civil
[191] Ibid.

you might love by sending you FREE samples from new and established brands."[192] They are employing a service that users are familiar with, but with more tangible offerings and with less transparency on the rationale for certain customers receiving certain products, "it's like Amazon's product recommendations, but real, so you can try, smell, feel, and taste the latest products. There is no obligation to purchase or review the product and you can opt out at any time."[193] Again, this practice seeks to maximize profits for the company, actually at the risk of data privacy and loss of consumer choice for customers, with "some of the products Amazon lists as being eligible for free samples include dog food, groceries, beauty products and health supplements; In other words, items you would buy repeatedly. It's easy to see why brands would pay to be a part of this program, especially if they are able to convince shoppers to go with their dog food or coffee over a brand they have been loyal to when shopping on Amazon in the past." [194] Sampling is an old and effective marketing technique because customers are more inclined to purchase a product more than once, and potentially become loyal to a company's entire product line and brand, if they are able to try a product once to see if they like it. However, this efficient path to sales may be met with concern because users have been opted in without Amazon disclosing this development publicly or with notification. The people that have been opted in and the products they receive are based on "Amazon's product recommendations". Customers do not necessarily understand that these recommendations are determined by machine learning that analyzes their personal data, or what that entails.

The concern that the artificial intelligence employed by companies to personalize ads will begin to know people better than they know themselves stems from the fear of machines and the resulting loss of control and sense of humanity. This sentiment is due to the superiority complex of the human species. The relinquishing of this superiority to machines causes people unease. Wendell Wallach and Collin Allen, authors of *Moral Machines,* do not believe that "increasing reliance on autonomous systems will undermine people's basic humanity...humans have always adapted to their technological products and the benefits to people of having autonomous machines around them will most likely outweigh the costs."[195] People may feel as though the nature of humanity is threatened because their choice of products is being influenced by machines. They feel as though the decisions for what they are exposed to or buy is not actually decisions they would have made if not for the use or intrusion of the machine. "We contend that some of the welfare-enhancing benefits of those technologies can backfire and generate consumer reactance if they undermine the sense of autonomy that consumers seek in their decision-making."[196] That may occur when consumers feel deprived of their ability to control their own choices: "predictive algorithms are getting better and better at anticipating consumers'

---

[192] Sherman, Erik. "Amazon Might Send You Samples Based on Every Purchase You Thought Was Private--Or Wanted to Forget." Inc.com. January 8, 2019.
https://www.inc.com/erik-sherman/amazons-new-secret-free-sample-program-could-out-creep-facebook.html

[193] Masters, Kiri. "Amazon Offers Product Sampling Program to Brands, Rooted in Machine Learning" Forbes, Jan, 2016.
https://www.forbes.com/sites/kirimasters/2019/01/09/amazon-offers-product-sampling-program-to-brands-rooted-in-machine-learning/#14378a7f3a56

[194] Sherman, Erik. "Amazon Might Send You Samples Based on Every Purchase You Thought Was Private--Or Wanted to Forget." Inc.com. January 8, 2019.

[195] Wendell Wallach and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*.

[196] André, Q., Carmon, Z., Wertenbroch, K. et al. "Consumer Choice and Autonomy in the Age of Artificial Intelligence and Big Data." Cust. Need. and Solut. (2018) 5: 28. https://doi.org/10.1007/s40547-017-0085-8

preferences, and decision-making aids are often too opaque for consumers to understand (how they might influence preferences and decisions)."[197] The point at which machines actually know what we want, or even know us better than we know ourselves, is unknown. The potential for that to happen, instills fear in many, and therefore creates an uneasiness around the use of targeted advertising. The central concern is that the use of machine learning in targeted advertising constitutes behavioral control, since it uses such personal information to influence individuals' purchasing decisions. If machine learning is ultimately making certain decisions *for* a customer through targeted advertising, it is (to an extent) hindering free choice. The use of the machine is not the only reason people to hesitate to embrace the use of artificial intelligence in advertising, how businesses and executives themselves are using this technology is also a main source of reticence.

83% of consumers believe personalized ads are morally wrong. This statistic highlights the already existing distrust of the use of machine learning to craft and deploy targeted ads to identified customers. This was discerned by RSA, a cybersecurity company, who conducted a survey of 6,000 adults in Europe and America that found that the majority of them believe companies using their data to personalize ads is unethical. This is mainly due to belief that their privacy is being violated, "due to constant media coverage, users are well-aware that technology's tracking of their behavior has been more pervasive than they assumed, and that their personal data has been shared with third (and sometimes fourth) parties in ways that feel violating. So it's no surprise that individuals are increasingly cynical about companies' data protection claims, promises, and policies."[198] about the utilization of personalization of ads so long as promises of privacy are upheld. The report has a definition for what ethical data use is, saying that 52% of all survey respondents agree that it is "when a company only takes the personal information needed to deliver the service customers are receiving and nothing more."[199] There are no ethical violations, however simply because the use of targeted ads may benefit businesses more than they benefit the customer. Consumers can view personalization as intrusive, however they cannot offer any rationale for why it is inherently immoral. Or perhaps it is not that they can't, but instead that the things that bother them are not at base moral or that it's no different morally from other forms of advertising. The majority of consumers do not like advertising and feel misled by ads that don't appear as explicitly as ads as they would on TV. However, this does not mean that it is morally wrong, simply a change to the advertising industry and methods that consumers will need to adapt to while also, ironically, benefitting from.

Advertising is typically seen as unethical because it is the relinquishing of control to a company, or at the very least the felt manipulation on behalf of the business. The use of artificial intelligence in advertising is no different except this time control is being relinquished to an even less understood or trusted source: machines. With the right policies in place, however, these ethical issues can be mitigated, and the benefit to both customers and companies from the use of targeted advertising can be maximized. Targeted advertising allows for companies to provide consumers with more relevant and personal product suggestions. This makes the user experience more convenient, cost-efficient, and time-effective. The main idea is to significantly decrease the time that is taken to scroll and sort one's way through endless product choices.

---

[197] Ibid.

[198] RSA Data Privacy and Security Survey 2019.
https://www.rsa.com/content/dam/en/misc/rsa-data-privacy-and-security-survey-2019.pdf

[199] RSA pg. 11

Fraud and Falsehood

Deceptive advertising occurs when advertisers use false, misleading, or unproven information to support its claims for its products. When it is done with the intent to mislead the consumer as opposed to an honest mistake it can be a criminal act. Deceptive advertising violates the Lantham Act to prevent trademark infringement and unfair competition and the FTC regulates advertising through its truth-in-advertising laws.

Utilizing artificial intelligence to tailor ads to the needs and wants of customers is beneficial when the customer is not met with irrelevant messages and instead only with relevant and engaging content, but it may still raise some ethical concerns regarding the validity of certain claims. More clicks on an ad leads to increased traffic on a site, and this increased traffic means more interest in and presumably demand of a product. Publishing data on the number of clicks on an ad can entice advertisers to publish their ads on a site and encourage customers to buy products shown. However, when these numbers are inflated by ad-clicks from computer-based bots or fake social network accounts that engage through likes and comments, the data is inherently false and misleading to the consumer. Perhaps more concerning, this inflation of data can actually lead to increased prices for consumers to a false but perceived increase in demand for a product.

Similar concerns of deceptive advertising practices, arise with the use of affiliate marketing. These companies essentially use targeted ads to attempt to gather customers that will most likely believe their pitches. Affiliate marketing's use of the artificial intelligence employed by Facebook is a practice that easily co-opts this useful technology to exploit customers through potentially misleading and false information about products purely to reap profits.

In the past, affiliates targeted people that they suspected would believe their pitches or fall for their schemes or by analyzing data looking for specific ages, geographic locations, or interests. Now, all of that analysis is done by Facebook's artificial intelligence. Because Facebook tracks who clicks on each ad, the advertisers can begin targeting everyone of similar demographics or people the algorithm determines will be likely to also purchase the product advertised. This practice has been realized and is ultimately successful: "affiliates describe watching their ad campaigns lose money for a few days as Facebook gathers data through trial and error, then seeing the sales take off exponentially."[200] It's not the artificial intelligence itself that is making immoral decisions but how people, businesses, and advertisers take advantage of it.

Robert Gryn, one of these so-called affiliate-marketers feels similarly about the ethics of affiliate marketing, but doesn't even believe the affiliates are to blame because as he says "they're just taking advantage of opportunities created by large corporations in a capitalistic system built around persuading people to buy things they don't need."[201] However, it wasn't until he began receiving handwritten complaints from people who had fallen victim to his deceptive sweepstakes for a free iPhone that resulted in recurring charges that he realized the detrimental impact it had on people. He was of the belief that because what he was doing was legal and feasible to make a fair amount of money that he could continue to do so with ease. It wasn't until

[200] Faux, Zeke. "How Facebook Helps Shady Advertisers Pollute the Internet." Bloomberg.
[201] Ibid.

he was met with the emotional harm that it caused that he thought about the faces and financial situations of the people, many of them among the poorest, behind the orders and numbers. It's not the artificial intelligence that makes advertising morally worse, it's that it makes advertisers more removed from the harm some of their strategies can cause. It's not the tool, it's the ease with which people can use the tool for broad negative impact without feeling as personally responsible for what they are doing.

Policy Solutions

There are many benefits to the use of artificial intelligence in advertising, mainly the cost-effectiveness and efficiency with which personalized content can be seen by a user. Additionally, the ads allow for the internet to remain free and open to all. This is due to the revenue generated by websites such as Facebook and Google that take in billion dollars annually from advertisers.[202] Ads can provide relevant informative content to users that want to know details about products they want or that they can discover. Algorithms can promote this relevancy which in turn creates this value for the ads.

Artificial intelligence has the capacity to leverage data to transform the customer experience, but at the potential cost of ethics and privacy. One of the main ways to mitigate this risk and keep the technology ethical is to be as transparent as possible according to Rob High, vice president and chief technology officer of IBM Watson.[203] So long as Facebook and Google adhere to and continue to improve upon their privacy policies in order to ensure the protection of users' identities, then the potential for violations in equality, freedom, and those resulting from deceptive advertising should be mitigated. However, they should actively seek to enforce their policies for all of their subsidiaries and proactively revise their policies to include legal as well as ethical challenges they may face.

This also means that because these types of policies involving artificial intelligence have already been implemented, employed, and modified, companies such as Facebook, Google, and Amazon should be models or even exemplars of what it means to ethically utilize this new technology. Because companies such as Facebook and Google are at the forefront of the use of such technologies they will be met with this burden, it will be essential in order to continue the practice's use so that other potentially smaller companies can benefit from such technologies to offset the inability to afford a large marketing and advertising budget. Targeted ads can also benefit both consumers and businesses that otherwise would not have been able to reach them. By promoting relevant content to users, the ads can create value and be an affordable option for smaller businesses that can actually match consumers' needs better than what is more readily visible. They can pay for their ad to be shown to the relevant people all through the mechanisms of machine learning.

It should not be up to private citizens and human rights organization to file litigation against discriminatory practices resulting from the use of ad tools. Facebook and Google should

---

[202] Ingram, Mathew. "How Google and Facebook Have Taken Over the Digital Ad Industry." Fortune. January 4, 2017. http://fortune.com/2017/01/04/google-facebook-ad-industry/

[203] Morgan, Blake. "Ethics and Artificial Intelligence with IBM's Rob High." Forbes. (https://www.forbes.com/sites/blakemorgan/2017/06/13/ethics-and-artificial-intelligence-with-ibm-watsons-rob-high/)

be proactively working to improve their technology so that users are actually protected against abuse. Rather than being reactive, companies should seek out ads policies that are proactive with the main goal of maintaining the highest ethical standards to protect data privacy, a strategy that will benefit them, customers, and the companies that advertise on their site as a result of a trusting relationship between the three entities. Other industries should be able to look at advertising privacy policies and shape their own policy to embody the values of these companies that already employ them.

Works Cited

AI for Advertising. IBM. https://www.ibm.com/watson-advertising

André, Q., Carmon, Z., Wertenbroch, K. et al. "Consumer Choice and Autonomy in the Age of Artificial Intelligence and Big Data." Cust. Need. and Solut. (2018) 5: 28. https://doi.org/10.1007/s40547-017-0085-8

Armstrong, Gary, Marketing: An Introduction 13th Edition.

Barton, Kris. "AI is within reach for small business marketing." Entrepreneur, Jan. 2019. https://www.entrepreneur.com/article/325817

Datta, A., M. C. Tschantz, and A. Datta (2015). "Automated Experiments on Ad Privacy Settings. Proceedings on Privacy Enhancing Technologies." 2015(1), 92–112. https://www.nber.org/chapters/c14011.pdf

Duhigg, Charles. "How Companies Learn Your Secrets." NY Times. February, 16, 2012. https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=1&hp

Faux, Zeke. "How Facebook Helps Shady Advertisers Pollute the Internet." Bloomberg. https://www.bloomberg.com/news/features/2018-03-27/ad-scammers-need-suckers-and-acebook-helps-find-them

Friedler, Sorelle A. and Wilson, Christo. "Potential for Discrimination in Online Targeted Advertising." Proceedings of Machine Learning Research 81: 1-15, 2018. http://proceedings.mlr.press/v81/speicher18a/speicher18a.pdf

Hill, Kashmir. "How Target Figured Out a Teen Girl Was Pregnant Before Her Father Did." Forbes. February 16, 2012. https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#1d3ac9136668

Ingram, Mathew. "How Google and Facebook Have Taken Over the Digital Ad Industry." Fortune. January 4, 2017. http://fortune.com/2017/01/04/google-facebook-ad-industry/

Jercinovic, Jason. "The Ethics of Using AI in Advertising." AdAge. July, 2017. https://adage.com/article/digitalnext/ethics-ai-advertising/309535/

Leathern, Rob. "New Ads Policy: Improving Integrity and Security of Financial Product and Services Ads." Facebook Business. January 30, 2018. https://www.facebook.com/business/news/new-ads-policy-improving-integrity-and-security-of-financial-product-and-services-ads

Masters, Kiri. “Amazon Offers Product Sampling Program to Brands, Rooted in Machine Learning.” Forbes, Jan, 2016. https://www.forbes.com/sites/kirimasters/2019/01/09/amazon-offers-product-sampling-program-to-brands-rooted-in-machine-learning/#14378a7f3a56

Martin, D. Kelly and Smith, N. Craig. “Commercializing Social Interaction: The Ethics of Stealth Marketing.” *Journal of Public Policy & Marketing*, Vol. 27, No. 1 (SPRING 2008), pp. 45-56. Sage Publications Inc.

Mittelstadt, Brent Daniel, et al. “The Ethics of Algorithms: Mapping the Debate.” Big Data & Society, Dec. 2016, doi:10.1177/2053951716679679.

Morgan, Blake. “Ethics and Artificial Intelligence with IBM’s Rob High.” Forbes. (https://www.forbes.com/sites/blakemorgan/2017/06/13/ethics-and-artificial-intelligence-with-ibm-watsons-rob-high/)

O’Neil, C. (2017). “Weapons of math destruction: How big data increases inequality and threatens democracy.” Broadway Books.

Parris, Terry and Angwin, Julia. “Facebook Lets Advertisers Exclude Users by Race.” ProPublica. https: //www.propublica.org/article/facebooklets-advertisers-exclude-users-by-race, 2016.

Reviewing Targeting to Ensure Advertising Is Safe and Civil. Facebook Business. https://www.facebook.com/business/news/reviewing-targeting-to-ensure-advertising-is-safe-and-civil

RSA Data Privacy and Security Survey 2019. https://www.rsa.com/content/dam/en/misc/rsa-data-privacy-and-security-survey-2019.pdf

“Summary of Settlements Between Civil Rights Advocates and Facebook.” ACLU. March 29, 2019. https://www.aclu.org/other/summary-settlements-between-civil-rights-advocates-and-facebook

Sherman, Erik. “Amazon Might Send You Samples Based on Every Purchase You Thought Was Private -- Or Wanted to Forget.” Inc.com. January 8, 2019. https://www.inc.com/erik-sherman/amazons-new-secret-free-sample-program-could-out-creep-facebook.html

Tucker, Catherine. “Privacy, Algorithms, and Artificial Intelligence.”

https://www.nber.org/chapters/c14011.pdf

Turner, John. “The Planning of Guaranteed Targeting Display Advertising.” Vol. 60, No. 1, January–February 2012, pp. 18–33. Operations Research.

Wendell Wallach and Colin Allen. “Moral Machines: Teaching Robots Right from Wrong.” *New York: Oxford University Press*, 2009

Winter, Jenifer. “Introduction to the Special Issue: Digital Inequalities and Discrimination in the Big Data Era.” *Journal of Information Policy,* Vol. 8 (2018), pp. 1-4. Penn State University Press.

Applications of AI in the Financial Services Sector

Coleman Kraemer

This chapter will cover the use and application of artificial intelligence and advanced machine learning in the broad field of finance. Among its many uses, artificial intelligence is used by lenders to calculate and better understand credit scores, banks to detect fraud and market manipulation, and hedge funds to shape trading strategies. There are many ethical implications and possible benefits and harms that come from the application of artificial intelligence in the world of finance, such as an increase in the accessibility of low-cost financial advising services or an increase in information asymmetry. While AI is used in many areas in financial services, this chapter will examine its application in the wealth management industry and in hedge fund trading strategies.

The Rise of Robo-Advisors

Consider the case of John, a college graduate who has $10,000 sitting in a bank account. He's about to start work in a new city and isn't sure what to do with this money. He doesn't need the money anytime soon; rather, he wants to invest and grow his wealth. One option is he could let it sit in a bank account and gain a minimal amount of interest. However, the purchasing power of this cash would decrease as time went on due to inflation. He thinks it's a good idea to invest in the stock market, but isn't really sure how to do so. He is also uninformed when it comes to financial planning and saving. He could open up a brokerage account and invest randomly in stocks, but that would likely just result in his losing money. Instead, he reads about these new robo-advisors that create a bespoke portfolio and invest your money for you for at an affordable price. John decides to sign up online— the next day all his assets are invested into a well-balanced portfolio. He no longer needs to worry about what to do with this money and can now focus on his career, without constantly monitoring the positions in his portfolio. These types of advisors are the future of wealth management advisory services.

Robo-advisors are a category of online financial services that provide financial advice and investment management solutions with minimal human involvement. This type of advisory service emerged out of the 2008 financial crisis, stemming from a societal distrust in more traditional financial services, particularly large banking institutions. Considered a breakthrough in traditional wealth management services, robo-advisors are artificial intelligence-driven virtual advisors that provide expert advice based on algorithms that aggregate a variety of considerations, such as risk tolerance, income, and age. According to the Concise Fintech Compendium, a traditional robo-advisor is defined as "a self-guided online wealth management service that provides automated investment advice at low costs and low account minimums,[204] employing portfolio management algorithms."[205] Companies that employ this type of advisory, such as Betterment, Wealthfront, and Wealthsimple, essentially act as low-fee wealth managers; while the typical wealth manager charges 1.0% to 1.25% of assets under management (AUM),

[204] An account minimum is the minimum amount of money needed in order to open up an account
[205] Schueffel, Patrick. *The Concise Fintech Compendium*. School of Management Fribourg, 2017, pp. 26.

these robo-advisors only charge 0.25% to 0.45%, depending on allocation preferences.[206] Over time, this difference in fees amount to a significant effect on long-term capital growth. For instance, in a twenty-year period, a 1.00% fee on AUM reduces the portfolio value around 15% compared to a fee of 0.25%. If your principal investment amount was $100,000, this would mean a reduction of $30,000 by the end of 20 years.[207]

These robotic advisors eliminate the need for a human financial advisor by delivering their customers direct software access to their investment portfolios at any time, coupled with a fully automated and easy-to-use help center. Due to a distrust in the traditional financial services sector after the Great Recession, their low management fees, and their user-friendly platforms, these financial technology companies have created a remarkable business plan that employs the use of artificial intelligence with a user risk-specific investment style.

Robo-advisors traditionally employ a passive and factor-based investment style. Passive investing is an investment strategy that seeks to maximize returns by minimizing the purchasing and selling of securities, focusing on the long-term purchase of index funds,[208] such as the S&P 500 or the Dow Jones Industrial Average.[209] The strategy of these robo-advisors utilizes Modern Portfolio Theory,[210] a strategy on how risk-averse investors can maximize returns based on a given amount of market risk. In other words, this strategy determines the correct balance of stocks, cash, and bonds to optimize the returns of a client's portfolio based on their specific risk profile.

These robo-advisors use software to choose index ETFs (Exchange-Traded Funds)[211] for their customers' given portfolios according to a risk desired by each particular client. ETFs are the building blocks for robo-advising platforms, invested in to gain exposure to certain market indices, geographies, or sectors, such as healthcare and technology.[212] Essentially, ETFs allow advisors to invest in a broader range of equities and securities, thus limiting their exposure to a single company. In terms of a customer's risk tolerance, these robo-advisors determine this tolerance from a questionnaire that examines liquidity factors and risk appetite. This questionnaire will ask questions like, "When deciding how to invest your money, what do you care about more: Maximizing gains, minimizing losses, or both equally?" or "The global stock market is often volatile. If your entire investment portfolio lost 10% of its value in a month during a market decline, what would you do?". These types of questions help assess how aggressive you are as an investor and how risk-averse you are.[213]

A higher risk tolerance equates to a higher percentage of one's portfolio in riskier assets, such as US small and midcap stocks or a higher allocation in Emerging Markets equities. Lower

---

[206] This refers to the amount that is allocated to different ETF funds that charge an addition percentage of AUM.
[207] "How Fees and Expenses Affect Your Investment Portfolio." *SEC*, www.sec.gov/investor/alerts/ib_fees_expenses.pdf.
[208] An index fund is a type of mutual fund that attempts to match a financial index, such as the S&P 500, Down Jones, or Nikkei.
[209] "Passive Equity Investing." *CFA Institute*, www.cfainstitute.org/membership/professional-development/refresher-readings/2019/passive-equity-investing.
[210] "What Is Wealthfront's Investment Strategy?" *Wealthfront*, support.wealthfront.com/hc/en-us/articles/115000562903-What-is-Wealthfront-s-investment-strategy-.
[211] An Exchange-Traded Fund is a marketable security that tracks a stock index or a basket of assets.
[212] Crager, Bill, and Jay W. Hummel. *The Essential Advisor: Building Value in the Investor-Advisor Relationship*. Wiley, 2016, pp.12.
[213] "Risk Profile Questionnaire." *Wealthfront*, www.wealthfront.com/questionnaire.

risk tolerance means a higher allocation to fixed income securities, such as municipal bonds and Treasuries. Through the use of artificial intelligence, these companies are able to study complex datasets and, in turn, cost-effectively deliver financial advisory solutions that are specifically tailored to meet the needs of a given client.[214] The service also uses artificial intelligence to track account activity in order to better deliver more tailored and specific financial advice.

This technology offers a cheap solution to the complexities of financial planning, which is usually defined by high fees and large allocation minimums; many wealth management firms usually require minimums that range from several hundred thousand to a million dollars. While more typical asset and wealth managers have a wealth minimum, companies like Betterment and Wealthfront either require incredibly low minimums or no minimum at all. Thus, mom and pop investors are able to allocate their money. Due to this lack of restrictions, people of lower incomes can access the type of financial planning services that are traditionally only available to the wealthy individuals with high net-worth. Robo-advisors solve the challenge of providing financial advice and solutions to a wider client base, particularly those that are less affluent and do not have a high level of investable assets.[215] Overall, these types of robo-advisor startups target a generally younger audience of millennials, who are traditionally more comfortable with an online platform. However, they have recently begun targeting a wider client base, particularly retirees who want to save money on their portfolio management fees.[216]

Some of these independent robo-advising companies, such as Betterment, offer a hybrid model that allows one to receive advice from human advisors at a small fee. At the moment, the allocation of the world's assets in robo-advisors is relatively small—the top four AI advisors in 2017 managed a total of $128 billion in assets.[217] However, it is estimated that by 2020, between $2.2 to $3.7 trillion dollars will be managed with the support of robo-advisory companies.[218] Large firms, like Fidelity, Charles Schwab, and Morgan Stanley, have been aggressively building their own robotic financial advisory branches. Some experts believe that every major bank will start to incorporate this type of advisory into their wealth management operations, thereby leading to a stark rise in competition for these types of services. Unfortunately, some see the long-term viability and survival potential of these companies as low due to the entry of these larger banks and wealth management firms. In other words, due to the success of these low-fee AI platforms, many traditional wealth management firms have begun to create their own robo-advising platforms, which will result in a higher amount of competition for companies like Betterment and Wealthfront.

While many robo-advisors share many similarities, such as low fees, adjustable portfolios, and an online advising platform, they do display a variety of differences. For many investors, the company you choose can represent the type of investor you are or the particular investment style. Companies like that have no minimum investments, such as Betterment, are designed for the less advanced investor with lower amounts of investable assets. Bloom is

---

[214] North America. "The Rise of the Robo-Advisor: How Fintech Is Disrupting Retirement." *Knowledge@Wharton*, 2018, knowledge.wharton.upenn.edu/article/rise-robo-advisor-fintech-disrupting-retirement/.
[215] Crager and Hummel, *The Essential Advisor*, pp. 48.
[216] Eisenberg, Richard. "Robo-Advisers: Not Just For Millennials Anymore?" *Forbes*, 6 Dec. 2016, www.forbes.com/sites/nextavenue/2016/12/06/robo-advisers-not-just-for-millennials-anymore/#299e28fa2738.
[217] "The Evolution of Robo-Advisors and Advisor 2.0 Model." *EY*, 2018, The evolution of Robo-advisors and Advisor 2.0 model.
[218] "The Expansion of Robo-Advisory in Wealth Management ." *Deloitte*, www2.deloitte.com/content/dam/Deloitte/de/Documents/financial-services/Deloitte-Robo-safe.pdf, pp.1.

designed specifically for people to invest through their employment 401K plans.[219] WiseBanyan is for those who hate fees that eat long-term portfolio gains—this company does not charge any fees to manage your account. Swell is advertised to those who want to be socially responsible investors, targeting companies that focus on renewable energy and producing zero waste.[220]

Ethical Implications

Removing the human component from financial advisory has a significant number of benefits, such as increasing accessibility to financial advice for the general public through lower portfolio management fees and minimums. The less affluent certainly benefit, as these AI advisors give expert financial guidance and allocate their client's investable assets for an affordable price. It also eliminates human fallibility and increases overall transparency between the client and advisor.[221] These automated advisors increase the transparency in portfolio management and fees through easy-to-use websites, platforms, and investment updates. Virtual advisors update their clients with any change in their allocation, a feature that is not usually included in most human financial advisor programs.[222] This transformation will also reduce the amount of violated regulations and laws; these types of artificial intelligence systems excel at adhering to compliance because such compliance is a part of their programming. The idea of outsourcing perfectly ethical behavior in the financial sector to AI is certainly attractive from an ethical perspective. Ultimately, given that AI advisors both extend the benefits of investing to a broader, underserved population and offer automatic, programmed compliance, there are obvious positive ethical implication for the rise of these AI advisors.

At the same time, there are certain clear ethical drawbacks and limitations to a fully automated financial advisor. At their current level, such automated investment platforms, while a certain black-and-white, by-the-book compliance, lack the depth of ethical judgment of real human advisors.[223] While one can certainly debate the level of ethical action in human beings, many financial advisors are registered fiduciaries, which means that the advisor is legally required to act in the best financial interest of his or her client. However, these artificial intelligence systems are not held to this standard. Not only are they unable to compute in an empathetic way due to the current limitations in programming artificial intelligence, these robo-advisors are also incapable of giving clients ethical advice. For instance, these advisors will be unable to advise against a certain high-risk tolerance in a client's portfolio, even if their client appears to prefer high-risk according to their survey or if the client carries high levels of debt. They are also unable to give nuanced advice to a client who is spending beyond his means; AI

---

[219] Hicks, Coryanne. "The Best Robo Advisor for Each Type of Investor." *U.S. News & World Report*, U.S. News & World Report, money.usnews.com/investing/investing-101/slideshows/the-best-robo-advisor-for-different-types-of-investors?onepage.

[220] Ibid.

[221] McCarty, Steven R. "Human Advisors vs. Robo-Advisors: Will Ethics Trump Compliance?" *ThinkAdvisor*, 17 Apr. 2015, www.thinkadvisor.com/2015/04/17/human-advisors-vs-robo-advisors-will-ethics-trump/?slreturn=20190302233127.

[222] Lewis, Michael. "The Rise of Virtual Financial Robo-Advisors for Your Investments - Types, Pros & Cons." *Money Crashers*, 17 Jan. 2019, www.moneycrashers.com/virtual-financial-robo-advisors-investments/.

[223] Beilfuss, Lisa. "The Future Robo Adviser: Smart and Ethical?" *The Wall Street Journal*, Dow Jones & Company, 20 June 2018, www.wsj.com/articles/the-future-robo-adviser-smart-and-ethical-1529460240.

advisors are unable to conduct subtle, human conversations that suggest to the client to save more or to stop living extravagantly. They are also unable to advise against potentially unethical securities, such as commercial gun manufacturers or questionable foreign entities. Thus, these robotic advisors are unable to guide their clients in a way a human advisor would and unable to help them consider all options and ethical nuances of their financial portfolio in order to maximize not only security for themselves and their family but also ethical clarity. While these robo-advisors cannot currently employ these types of fiduciary and thoughtful services, there is much hope that over time they will be able to do so. If this software was able to employ AI that could think at a fiduciary level, then this system of wealth management would far surpass the usefulness and need of human wealth management advisors.

Actual portfolio performance is a key ethical issue for the rise of these robotic advisors. While this type of passive investing has done relatively well in the past ten years, one question remains: How well will these types of artificial intelligence portfolios behave when faced with a recession or market crash? These platforms are able to rebalance their portfolio and harvest losses more efficiently than human advisors. They also do not fall prey to human and investor bias, limiting the possibility of losses due to irrational market fears. However, since these robo-advisors do not have a track record of performance in more volatile markets, it is difficult to tell how this passive investment style will do under market-downturn conditions compared to a more human controlled active investment style.

These services have also been known to over-emphasize reliance on client risk tolerance questionnaires. Since they do not have the intuitive judgment of in-person human advisors, these AI advisors have to rely solely on online questionnaires to determine the risk tolerance of their clients, which can be uninformative or misleading since many do not check certain crucial information, such as credit card history and number of dependents. These questionnaires are also self-administered, which often results in biased and unrepresentative answers.[224] Ultimately, many have seen these questionnaires as vague and imprecise assessment tools, severely limiting the overall level of advice and guidance these advisors can give.[225]

The many benefits of robo-advisory services greatly outweigh the negatives. While much work has to be done to update the AI component of these advisors in order to account for fiduciary responsibilities and more concrete and detailed risk-profile questionnaires, these advisors provide many societal benefits, particularly to those who are less affluent. This application of artificial intelligence allows people with lower income to access financial advice and customized portfolios that would not be available to them at traditional wealth management firms. That said, only if these AI advisors prove effective in the long run—particularly through an inevitable downturn in the market—will extending their use to a broader underserved population make them ethical.

Trading Strategies

[224] Sironi, Paolo. *FinTech Innovation: From Robo-Advisors to Goals Based Investing and Gamification*. Wiley Financial Series, 2016, pp. 47.
[225] Ji, pp.1565.

On April 23, 2013, the Associated Press tweeted a shocking alert: "Breaking: Two Explosions in the White House and Barack Obama is injured."[226] Unsurprisingly, this post caused some drastic initial reactions, including a sudden crash of the financial markets, wiping out $130 billion in shareholder value, with liquidity collapsing as well. This news alert was, of course, a false twitter post, attributed to Syrian hackers, which became evident after five minutes, with financial markets eventually rebounding. This event was not the only time computer algorithms and artificial intelligence trading systems have caused an immediate reversal in market prices, such as the infamous 2010 Flash Crash. These occurrences represent the significant power computer algorithms hold on the financial market. Due to their lightning quick ability to buy and sell securities on a gigantic scale, these computer systems have the ability to crash market prices and dry up liquidity, potentially causing significant, lasting damage.

The incorporation of artificial intelligence and machine learning in financial trading strategies has increased significantly over the past decade, particularly in hedge funds, asset managers, and financial institutions. This reliance on technology and machine learning grew out of the belief that human emotions negatively affect the performance of a portfolio. In other words, people can be too emotionally tied to their investments, losses, and gains, which can cause them to often make mistakes. Many asset managers and broker-dealers allow these artificial intelligence systems to trade without human oversight, allowing these computers to make thousands of trades a day with little, if any, supervision. These managers believe that this technology holds significant advantages over traditional trading and analysis, allowing them to generate alpha, or the ability for a manager or a strategy to outperform the general financial market. However, it should be emphasized that many artificial intelligence algorithms turn out to be ineffective and useless. AI is not magical; it is only as good as its dataset and software engineers.

These artificial intelligence trading systems analyze enormous datasets at extraordinary speeds that are vastly superior to human cognition. AI trading software can absorb enormous volumes of data to predict financial markets. This technology can quickly analyze wide ranging information, from social media posts and news articles to equity research reports and consumer data. These systems are continuously learning through the information from this analyzed data, which then decide on the "best" executable trades.[227] AI securities trading is categorized under algorithmic trading, which uses powerful computers to run complex mathematical formulas to execute trades and make decisions.[228] Algorithmic trading allows a firm to make complex high frequency trades, allowing the trader and system to make thousands of trades per second.

There are a wide variety of ways in which artificial intelligence is used in specific trading strategies. For instance, it is employed in trade execution algorithms that use a Volume Weighted Average Pricing strategy, breaking up orders into smaller ones to minimize the impact on equity pricing.[229] There are also algorithms that take advantage of larger price movements that come with a significant order size.[230] AI uses historical financial market data to test portfolios on a

---

[226] Baumann, Nick. "How One Tweet Almost Broke US Financial Markets." *Mother Jones*, 25 April 2013, www.motherjones.com/politics/2013/04/associated-press-hacked-tweet-high-speed-trading/.
[227] Bernard, "The Revolutionary Way of Using Artificial Intelligence in Hedge Funds."
[228] Chen, James. "Algorithmic Trading Definition." *Investopedia*, 27 Feb. 2019, www.investopedia.com/terms/a/algorithmictrading.asp.
[229] "Machine Learning for Trading." *Sigmoidal*, 15 Oct. 2018, sigmoidal.io/machine-learning-for-trading/.
[230] Ibid.

wide scale of negative and positive scenarios, such as a stock market bubble or a housing crisis. Artificial intelligence can also be used to conform to compliance regulations, helping firms meet pre and post-trade regulatory requirements.[231]

While a variety of asset managers use artificial intelligence to execute trades, the most revolutionary types of this technology are being used in hedge funds.[232] Today, there are more than ten thousand hedge funds managing over $3 trillion in assets.[233] And, according to a Barclays PLC survey, approximately 62% of hedge funds use artificial intelligence technology in some way to inform their trading strategies and market outlook.[234] These AI applications include risk management, portfolio construction, idea generation, and trade execution. Therefore, there is a large amount of capital that is being run through artificial intelligence trading systems in hedge funds every day. Hedge funds that rely on artificial intelligence and computer algorithms are known as quantitative hedge funds, which have a reputation of performing above their peers and the market. These hedge funds grew from algorithmic trading to incorporate machine learning and deep learning to drive trade decisions. Some of the largest hedge funds employ these types of strategies, included AQR Capital Management, Renaissance Technology, Two Sigma, Bridgewater Associates, and Point72 Asset Management.

Storing the significant amount of data in order for an AI system to work correctly and continue learning is incredibly expensive, which means only certain types of institutions can use them. Hedge funds employ the most complex and sophisticated forms of artificial intelligence algorithms due to their high fee structures, large capital allocations, and high performance expectations. A significant amount of work has been done using machine learning to predict the direction of asset pricing, which is the examining of historical pricing and trading volumes to forecast future values.[235] Human "quants" usually accompany or monitor these AI trading systems, who fine tune the algorithms in real-time as new scenarios present themselves. These asset managers also use them to improve efficiency in their middle and back office operations, such as accounting and investor relations.[236]

Ethical Implications

Due to the great sensitivity of AI systems, the concomitant volatility of a market influenced by AI algorithms is itself an area of ethical concern. As the story at the beginning of this section illustrated, the application of AI high-frequency algorithms can cause significant

---

[231] "Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications." *Financial Stability Board*, 2017, www.fsb.org/wp-content/uploads/P011117.pdf, pp.18.

[232] A hedge fund is an investment vehicle that is known to use high risk investment methods in order to garner larger investment gains. These funds are known to charge very high fees for performance and AUM.

[233] Marr, Bernard. "The Revolutionary Way of Using Artificial Intelligence in Hedge Funds." *Forbes*, Forbes Magazine, 6 Mar. 2019, www.forbes.com/sites/bernardmarr/2019/02/15/the-revolutionary-way-of-using-artificial-intelligence-in-hedge-funds-the-case-of-aidyia/#751de7f157ca.

[234] Dravis, Paul. "Artificial Intelligence in Finance: The Road Ahead." *Future Perfect Machine*, May 2018, pp.9

[235] Choudhry, Rohit, and Kumkum Garg. "A Hybrid Machine Learning System for Stock Market Forecasting" *World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering*, vol. 2, no. 3, 2008, pp. 689.

[236] Salvage, Peter. "Artificial Intelligence Sweeps Hedge Funds." *BNY Mellon*, Mar. 2019, www.bnymellon.com/us/en/what-we-do/business-insights/artificial-intelligence-sweeps-hedge-funds.jsp.

negative effects to the financial markets. While these effects have been generally short-lived, their effects can greatly impact short-term derivatives and overall market liquidity and sentiment. There are also instances of AI systems making small errors that cost respective hedge fund hundreds of millions of dollars. For example, in August 2012, the AI systems at Knight Capital Group mistakenly kept submitting the same equity orders to the SEC, which resulted in the execution of over four million transactions and billions of dollars of exposure in both long and short positions. This led to $460 million in trading losses and a share price collapse of its own stock of 75%.[237] Rightfully so, the SEC later sanctioned Knight Capital Group for this incident, claiming that it failed to meet risk management procedures.[238] In the end, these artificial intelligence trading systems hold tremendous power over their capital and the financial markets. Therefore, due to this enormous risk and influence, the hedge funds and other asset managers should be held accountable for the outcomes of these AI technologies.

These examples of artificial intelligence use in hedge funds are some of the more complex applications of the technology due to the immense potential monetary reward. Ultimately, these complexities add to the uncertainty of this technology. There are some cases where these technologies become so complex that not even the people overseeing them understand their decisions and deep learning adaptive properties. These self-altering systems can even be misunderstood by their own creators. For instance, Man Group, a hedge fund that manages approximately $96 billion dollars, realized that its AI trading program was beginning to execute trades that its software engineers couldn't even understand.[239] Due to the problems that this gap in understanding could cause in the marketplace, there should be careful regulatory scrutiny over these "black box" hedge funds, to make sure that these experimental AI systems do not cause future financial meltdowns. The people in charge and operating these machines must remain in complete control of these systems and not allow them to gain complete autonomy. There have been instances were complex asset managers have cause tremendous damage on the financial markets and economy, such as the collapse of Long-Term Capital Management that nearly caused a financial catastrophe due to the incredible lack of regulatory oversight and highly leveraged and complex positions. Thus, these systems should be heavily regulated by financial authorities.

Clearly, there should be extensive safeguards placed in these trading technologies—regulators should not let history repeat itself in this rise of this new technology. However, this regulation will be incredibly difficult due to the sheer complexities of these systems. Unfortunately, it is likely that financial markets regulators do not have the knowledge or expertise to regulate artificial intelligence trading effectively.[240] Thus, governmental regulators need to place more attention on funding this type of regulation, employing the many software engineers who actually understand the technology to create safeguards to prevent catastrophic market events.

---

[237] Wellman, Michael, and Uday Rajan. *Ethical Issues for Autonomous Trading Agents*. strategicreasoning.org/wp-content/uploads/2017/01/ethical-issues-autonomous.pdf, pp.3.
[238] Ibid., 3.
[239] Satariano, Adam. "The Massive Hedge Fund Betting on AI." *Bloomberg*, Sept. 2017, www.bloomberg.com/news/features/2017-09-27/the-massive-hedge-fund-betting-on-ai.
[240] Nicholls, Marloes. *Briefing: Ethical Use of AI*. The Finance Innovation Lab, Feb. 2018, financeinnovationlab.org/wp-content/uploads/2018/04/Briefing-Ethical-Use-of-AI-in-Finance.pdf, pp.6.

The most concerning effect of this application of artificial intelligence is information asymmetry. These AI platforms are able trade and comprehend information that is vastly superior to human traders and regular, less advantaged investors. Thus, those that do not have high amounts of investable cash and net worth will not be able to allocate any money to these funds, unable to partake in their profits. This application of AI will ultimately lead to a decrease in the share of available information to the public, greatly favoring these quant funds over the rest of the investment community due to their complex technologies. However, that does not mean that this use of technology should be limited to hedge funds simply because it helps them gain an advantage. Instead, in addition to increased oversight, perhaps a tax policy should apply to hedge funds that are in this position, where they are taxed at a higher rate as their returns increase due to this AI technology.

The ethical implications of artificial intelligence in wealth management and trading strategies share some similarities. Both of these technologies have not been specifically tested in a time of a recession and a highly volatile market, so these uses both contain unquantifiable risk.[241] The looming question will be what these AI systems will do when an entirely unfamiliar situation comes and whether these systems will be able to truly learn as new, unforeseen events occur in the economy.

AI and Finance

Ultimately, whoever is in charge of the implementation of these systems, whether that be the regulators or asset managers, needs to recognize that artificial intelligences are not built with encoded moral guidance or ethics. While all the possible applications of AI in the financial services were not examined in this chapter, the examples of AI in robo-advisers and securities trading illustrate two different cases with varying ethical implications.

These uses come with their own range of ethical implications, which require certain safeguards and regulations to be implemented correctly. Robo-advisors represent a more positive application of artificial intelligence, which use AI-fueled technology to give advice to younger and less privileged investors, as well as passively managing their investment portfolios. This type of investment tool is incredibly important for those who would otherwise not have access to it. To be able to have an artificial intelligence advisor that creates a customized financial plan and allocates one's assets at an affordable rate is invaluable. It will be interesting what the long-term effects these robo-advisors have on wealth generation for the middle and lower classes.

Somewhat on the other end of the spectrum, quantitative hedge funds use these types of deep learning AI systems to gain an advantage on the global market, enabling them access to more information and the potential of higher returns. This presents quite an information asymmetry in the global markets, where only some types of managers will have access to invaluable information and be able to act quickly and efficiently with it. Clearly, increased regulation and scrutiny should be placed on this application of artificial intelligence, due to the possible consequences and negative effects these trading systems can have on the financial markets. Therefore, the use of AI is somewhat of a double-edged sword when considered in light

[241] Edwards, Helen, and Dave Edwards. "AI Does Not Have Enough Experience to Handle the next Market Crash." *Quartz*, Quartz, 12 Dec. 2017, qz.com/1151664/ai-does-not-have-enough-experience-to-handle-the-next-market-crash/.

of these two cases. While these AI advisors enable the less affluent to allocate their money efficiently, this positive is contradicted in sorts by AI's application in hedge fund strategies, which produce a stark information asymmetry between these classes of investors.

In the end, the effects of artificial intelligence mirror many of the other technological progresses that have befallen society. With the rise of technology comes the initial concern for its effect on people's daily lives, job security, and overall welfare. While these concerns are often over-emphasized, these technologies do hold noticeable risks for market destabilization and social inequalities that should be weighed when considering their ethical implications. However, these applications will likely end up having a positive impact on the financial services sector, as long as they are properly regulated, by increasing the amount of information being analyzed and the ability for cheaper financial planning and capital allocation.

Works Cited

"Artificial Intelligence and Machine Learning in Financial Services: Market Developments and Financial Stability Implications." *Financial Stability Board*, 2017, www.fsb.org/wp-content/uploads/P011117.pdf, pp.18.

Baumann, Nick. "How One Tweet Almost Broke US Financial Markets." *Mother Jones*, 25 April 2013, www.motherjones.com/politics/2013/04/associated-press-hacked-tweet-high-speed-trading/.

Beilfuss, Lisa. "The Future Robo Adviser: Smart and Ethical?" *The Wall Street Journal*, Dow Jones & Company, 20 June 2018, www.wsj.com/articles/the-future-robo-adviser-smart-and-ethical-1529460240.

Chen, James. "Algorithmic Trading Definition." *Investopedia*, 27 Feb. 2019, www.investopedia.com/terms/a/algorithmictrading.asp.

Crager, Bill, and Jay W. Hummel. *The Essential Advisor: Building Value in the Investor-Advisor Relationship*. Wiley, 2016.

Choudhry, Rohit, and Kumkum Garg. "A Hybrid Machine Learning System for Stock Market Forecasting" *World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering*, vol. 2, no. 3, 2008.

.

Dravis, Paul. "Artificial Intelligence in Finance: The Road Ahead." *Future Perfect Machine*, May 2018.

Eisenberg, Richard. "Robo-Advisers: Not Just For Millennials Anymore?" Forbes, 6 Dec. 2016, www.forbes.com/sites/nextavenue/2016/12/06/robo-advisers-not-just-for-millennials-anymore/#299e28fa2738

"The Evolution of Robo-Advisors and Advisor 2.0 Model." *EY*, 2018, The evolution of Robo-advisors and Advisor 2.0 model.

"The Expansion of Robo-Advisory in Wealth Management," *Deloitte*, www2.deloitte.com/content/dam/Deloitte/de/Documents/financial-services/Deloitte-Robo-safe.pdf.

Garbade, Michael J. "Clearing the Confusion: AI vs. Machine Learning vs. Deep Learning Differences." *Towards Data Science*, 14 Sept. 2018, towardsdatascience.com/clearing-the-confusion-ai-vs-machine-learning-vs-deep-learning-differences-fce69b21d5eb.

Hicks, Coryanne. “The Best Robo Advisor for Each Type of Investor.” *U.S. News & World Report*, U.S. News & World Report, money.usnews.com/investing/investing-101/slideshows/the-best-robo-advisor-for-different-types-of-investors?onepage.

“How Fees and Expenses Affect Your Investment Portfolio.” *SEC*, www.sec.gov/investor/alerts/ib_fees_expenses.pdf.

Ji, Megan. “Are Robots Good Fiduciaries? Regulating Robo-Advisors Under the Investment Advisers Act of 1940.” *Columbia Law Review*, 2017, doi:10.2139/ssrn.3036722

Lewis, Michael. “The Rise of Virtual Financial Robo-Advisors for Your Investments - Types, Pros & Cons.” *Money Crashers*, 17 Jan. 2019, www.moneycrashers.com/virtual-financial-robo-advisors-investments/.

“Machine Learning for Trading .” *Sigmoidal*, 15 Oct. 2018, sigmoidal.io/machine-learning-for-trading/.

Marr, Bernard. “The Revolutionary Way of Using Artificial Intelligence in Hedge Funds.” *Forbes*, Forbes Magazine, 6 Mar. 2019, www.forbes.com/sites/bernardmarr/2019/02/15/the-revolutionary-way-of-using-artificial-intelligence-in-hedge-funds-the-case-of-aidyia/#751de7f157ca.

McCarty, Steven R. “Human Advisors vs. Robo-Advisors: Will Ethics Trump Compliance?” *ThinkAdvisor*, 17 Apr. 2015, www.thinkadvisor.com/2015/04/17/human-advisors-vs-robo-advisors-will-ethics-trump/?slreturn=20190302233127.

Nicholls, Marloes. *Briefing: Ethical Use of AI*. The Finance Innovation Lab, Feb. 2018, financeinnovationlab.org/wp-content/uploads/2018/04/Briefing-Ethical-Use-of-AI-in-Finance.pdf

“Passive Equity Investing.” *CFA Institute*, www.cfainstitute.org/membership/professional-development/refresher-readings/2019/passive-equity-investing.

“The Rise of the Robo-Advisor: How Fintech Is Disrupting Retirement.” *Knowledge@Wharton*, 2018, knowledge.wharton.upenn.edu/article/rise-robo-advisor-fintech-disrupting-retirement/.

Sironi, Paolo. *FinTech Innovation: From Robo-Advisors to Goals Based Investing and Gamification*. Wiley Financial Series, 2016.

“Risk Profile Questionnaire.” *Wealthfront*, www.wealthfront.com/questionnaire.

Satariano, Adam. “The Massive Hedge Fund Betting on AI.” *Bloomberg*, Sept. 2017, www.bloomberg.com/news/features/2017-09-27/the-massive-hedge-fund-betting-on-ai.

Schueffel, Patrick. *The Concise Fintech Compendium*. School of Management Fribourg, 2017.

“What Is Factor Investing?” *BlackRock*, www.blackrock.com/us/individual/investment-ideas/what-is-factor-investing.

“What Is Wealthfront's Investment Strategy?” *Wealthfront*, support.wealthfront.com/hc/en-us/articles/115000562903-What-is-Wealthfront-s-investment-strategy-.

Salvage, Peter. “Artificial Intelligence Sweeps Hedge Funds.” *BNY Mellon*, Mar. 2019, www.bnymellon.com/us/en/what-we-do/business-insights/artificial-intelligence-sweeps-hedge-funds.jsp.

Wellman, Michael, and Uday Rajan. *Ethical Issues for Autonomous Trading Agents*. strategicreasoning.org/wp-content/uploads/2017/01/ethical-issues-autonomous.pdf, pp.3.

Artificial Intelligence and Healthcare

Shweta Lodha

Introduction

Artificial intelligence is revolutionizing almost all fields of work, and the healthcare industry is no exception. AI currently has applications in areas of early detection and diagnosis, treatment, outcome prediction, and prognosis evaluation. Major disease areas, including oncology, neurology, and cardiology are in the process of integrating AI tools to help inform better patient care and reduce diagnostic and therapeutic errors. Somashekhar et al. recently demonstrated that IBM Watson, an AI system developed by researchers at IBM, could reliably assist the diagnosis of cancer.[242] To accomplish this, researchers first trained Watson to interpret more than 600,000 pieces of medical evidence and two million pages from medical journals, allowing the AI system to develop an extensive breadth of knowledge about cancer treatment and diagnosis. In some cases, AI models have even shown greater accuracy in interpreting and diagnosing medical conditions than healthcare professionals. When used to diagnose lung cancer, Watson showed an accuracy rate of 90%, outperforming healthcare professionals, who performed with 50% diagnosis accuracy.[243] Furthermore, a paper recently published in Nature Digital Medicine reported the use of computer vision, a field that uses artificial intelligence models to automate tasks normally performed by the human visual system, to interpret echocardiograms at accuracies that were greater than those demonstrated by trained experts.[244,245] Together, these examples demonstrate that AI has the potential to enhance various facets of diagnosis and treatment.

The successful application of AI in healthcare is facilitated by the availability of large amounts of healthcare data. AI uses algorithms to learn from large volumes of healthcare data and subsequently provide clinical insights that are otherwise not easily gleaned. Before AI systems can be used, they must be "trained" using clinical data generated by the screening, diagnosis, and treatment assignment of a large number of patients. Specific types of clinical data used to create such AI include demographic patient information, medical notes, physical examinations, and clinical laboratory and images.[246] While it is clear that AI has the potential to enhance clinical decision making, the use of patient healthcare data to create effective AI technology raises several ethical questions. Given various biases inherent to the healthcare data used to create AI technologies, there may exist concerns regarding whether healthcare AI will

---

[242] Somashekhar, S. P., R. Kumarc, A. Rauthan, K. R. Arun, P. Patil, and Y. E. Ramya. "Abstract S6-07: Double Blinded Validation Study to Assess Performance of IBM Artificial Intelligence Platform, Watson for Oncology in Comparison with Manipal Multidisciplinary Tumour Board–First Study of 638 Breast Cancer Cases." (2017): S6-07.

[243] Steadman, Ian. "IBM's Watson Is Better at Diagnosing Cancer than Human Doctors." WIRED. October 04, 2017. Accessed April 20, 2019. https://www.wired.co.uk/article/ibm-watson-medical-doctor.

[244] Koch, Marta. "Artificial Intelligence is Becoming Natural." *Cell* 173, no. 3 (2018)

[245] Huang, T. "Computer Vision: Evolution and Promise" from CERN Document Server (1996).

[246] Huang, Ibid.

simply exacerbate existing inequities. Furthermore, given the important role patient data plays in generating effective AI, questions may also arise regarding the challenges healthcare AI might pose for patient confidentiality. Finally, existing uncertainties about the specific mechanisms through which AI systems make decisions might conflict with current standards for gaining informed consent.

This chapter will examine the ethical conflicts imposed by the use of biased data to generate effective AI systems and the impact AI healthcare models might have on the preservation of patients' rights to privacy and informed consent. Specifically, this chapter will discuss different sources of potential bias that can be encoded within and propagated by AI models used in healthcare. This chapter will also discuss possible conflicts imposed by healthcare AI on patients' rights to privacy and the informed consent process. This chapter will end with a series of policy recommendations that can help resolve some of the ethical conflicts highlighted within the chapter.

Bias

To create AI that enhances various facets of healthcare, programmers must train systems using a diverse compilation of data points. Examples of the type of data that AI health systems may require include but are not limited to patient trait inputs such as age, gender, disease history, and disease-specific data, like diagnostic imaging, gene expressions, electrophysiology test, physical examinations, clinical symptoms, and medications.[247] Bias may exist in both the type of data that the AI is trained with, and how the AI is programmed to interpret given data sets to generate outputs.

In order for AI to be effective in a healthcare setting, it must be trained. By effect, challenges arise when there are deficiencies in the training set and subsequent mismatches between training and operational data, or input data a model encounters once it has been trained. Limited data sets can result in a distributional shift, which occurs when previous experiences are not adequate for novel situations. Machine learning systems have been shown to be poor at recognizing changes in input data and may subsequently make erroneous predictions.[248] Recently, researchers Esteva et al. showed the potential for deep convolutional neural networks to classify skin lesions.[249] The researchers used pictures of lesions biopsied in a clinic to train the algorithm and stated in the article that their results may inform the use of such technology in mobile phones to make lesion classification services accessible to many people at once. However, since the system for diagnosing skin malignancy was trained with a different data set consisting of exclusively biopsied lesions, it may not perform as well when applied to the task of screening the general population, where the appearance of lesions and patients risk profile both

---

[247] Jiang, Fei, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. "Artificial Intelligence in Healthcare: Past, Present and Future." *Stroke and Vascular Neurology* 2, no. 4 (2017): 230-243.

[248] Challen, Robert, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. "Artificial Intelligence, Bias and Clinical Safety." *BMJ Qual Saf* 28, no. 3 (2019): 231-237.

[249] Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." *Nature* 542, no. 7639 (2017): 115.

vary.[250] It may be possible to address such a limitation by training models through transfer learning, a machine learning method whereby a model developed for a unique task is then reused as the starting point to train another model for a different task with similar attributes.[251] In this case, the model that was developed to recognize biopsied lesions can then be used as the starting point to train another model that can detect lesion images taken from a camera phone. Further research is necessary to determine the feasibility of this approach in decreasing incongruities between training and operational data sets of skin lesions.

Furthermore, disease patterns change over time, which can in turn intensify possible mismatches between training and operational data. For example, in a study by Davis et al., researchers explored the phenomena of prediction drift, in which over time, decreasing acute kidney injury incidence was associated with increasing false positives from the machine learning system. In other words, the machine learning system showed to be poor at recognizing relevant changes in the rate of acute kidney injury due to mismatches between training and operational data. This, in turn, resulted in the machine learning system over-predicting risk of acute kidney injury in patients.[252] Problems such as this may be resolved by retraining AI systems continuously or periodically. In some cases, it may be beneficial for AI health systems to incorporate inputs after a period of time instead of continuously to ensure stability and reproducibility of results.[253] For example, an AI system which continuously incorporates inputs and updates its model may produce inconsistent outputs during a certain period of time, which can impose confusion and organizational havoc upon healthcare providers and facilities at large. Alternatively, certain healthcare events might benefit from continuous retraining more than others. For example, it may be beneficial to use a continuously retraining AI model to diagnose and evaluate novel diseases that have not been widely studied. On the other hand, AI models used to diagnose disease states with easily recognizable and consistent phenotypes may be disrupted less by and subsequently benefit more from periodic retraining.

Biases may also be inherent to the medical datasets openly available for use by AI researchers. The vast majority of accessible healthcare data has been collected through clinical studies using white and male participants.[254] For example, a cross-sectional population based analysis of all patients enrolled in the National Cancer Institute's Clinical Trial Cooperative Group showed that men were more likely to enroll than women and Hispanic individuals were least likely to enroll. Furthermore, the data showed that patients enrolled were 24% less likely to be black.[255] As a result, treatment interventions have historically not been tailored to women and minority groups, yielding generally poorer treatment outcomes and longitudinal health outcomes

---

[250] Challen.

[251] Curry, Brian. "An Introduction to Transfer Learning in Machine Learning." *Medium*. July 26, 2018. Accessed April 11, 2019. https://medium.com/kansas-city-machine-learning-artificial-intelligen/an-introduction-to-transfer-learning-in-machine-learning-7efd104b6026.

[252] Challen.

[253] Anderson, Michael, and Susan Leigh Anderson. "How Should AI Be Developed, Validated, and Implemented in Patient Care?." *AMA Journal of Ethics* 21, no. 2 (2019): 125-130.

[254] Froomkin, A. Michael, Ian Kerr, and Joelle Pineau. "When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning." *Ariz. L. Rev.* 61 (2019): 33.

[255] Murthy, Vivek H., Harlan M. Krumholz, and Cary P. Gross. "Participation in Cancer Clinical Trials: Race-, Sex-, and Age-based Disparities." *JAMA* 291, no. 22 (2004): 2720-2726.

for these groups.[256] The use of such limited data sets to train AI health programs may recapitulate and even intensify long standing health disparities.[257] In order to get more representative data, it would be necessary to recruit female and minority physicians, as most minority patients seek physicians of their own race, and find out about clinical trials from their doctors. Furthermore, it may be necessary for sponsors of clinical trials to demonstrate the importance of the trial and potential benefit for patients and their communities. Specifically, sponsors should advocate for clinical trials to leaders of minority communities. While such solutions may be challenging to implement, multisector collaboration would likely improve the feasibility of getting more representative data.[258]

Training machine learning models with subjective and expressive clinical notes may also exacerbate bias. A study analyzed two case studies in which machine learning algorithms were used on 25,879 clinical and psychiatric patient stay notes to predict intensive care unit mortality and 30-day psychiatric readmission. Results demonstrated variance in prediction accuracy and subsequent machine bias with respect to gender and insurance type for ICU mortality and insurance policy for psychiatric 30-day readmission. Specifically, the machine learning algorithm had a greater error rate for predicting female ICU patients' mortality rates compared to those of male ICU patients. Furthermore, the algorithm had a greater error rate for predicting the mortality rates of ICU patients with public insurance compared to those with private insurance.[259]

These results may reflect how training algorithms with potentially biased clinical findings can in turn yield discriminatory and inaccurate outcomes. Of interest would be a parallel study assessing the error rates of health care providers predicting mortality rates of ICU patients varying in gender and insurance status. If health care providers yield greater error rates than did the AI model, it may still be permissible to use such a system in a healthcare setting to help predict ICU patients' probability of dying. A limitation of the study was that it did not report whether the errors made by the model were over predictions or under predictions of mortality rates.

AI used in healthcare may not solely show bias through use of limited and incomplete data sets. Algorithms can be created to organize the data in different ways that can then yield varied outcomes. For example, supervised learning relies on mapping inputs to known output labels, and is done using a ground truth, or prior knowledge of specific output values. On the other hand, unsupervised machine learning does not have labeled outputs, and instead generates outputs by detecting natural structures present within a data set.[260,261] Furthermore, there exists two major types of unsupervised learning methods, clustering and principal component analysis,

---

[256] Chen, Irene Y., Peter Szolovits, and Marzyeh Ghassemi. "Can AI Help Reduce Disparities in General Medical and Mental Health Care?." *AMA Journal of Ethics* 21, no. 2 (2019): 167-179.

[257] Gershgorn, Dave. "If AI Is Going to Be the World's Doctor, It Needs Better Textbooks." Prescription AI. September 10, 2018. Accessed April 21, 2019. https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks/.

[258] Coakley, Meghan, Emmanuel Olutayo Fadiran, L. Jo Parrish, Rachel A. Griffith, Eleanor Weiss, and Christine Carter. "Dialogues on Diversifying Clinical Trials: Successful Strategies for Engaging Women and Minorities in Clinical Trials." *Journal of Women's Health* 21, no. 7 (2012): 713-716.

[259] Chen.

[260] Jiang.

[261] Soni, Devin. "Supervised vs. Unsupervised Learning." Towards Data Science. March 22, 2018. Accessed April 11, 2019. https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d.

that find patterns within the data in different ways and subsequently can generate diverse outcomes.[262]

Different health problems can warrant the use of AI models developed through different learning methods. For example, the ability for supervised learning to predict known outputs based on certain inputs may make it useful for tasks such as classifying images of lesions, just as it is used for handwriting recognition. Supervised learning can also be used to classify documents, such as distinguishing between a clinical trial discussing renal failure and a systematic review about the mechanisms motivating kidney disease development. Supervised learning may also be useful for classifying patients by risk.[263] However, deciding the specific risk factors that the model should use to rank patients may introduce bias. For example, there may be uncertainty regarding whether the AI should take into consideration factors such as the insurance status of the patient, or use a utilitarian approach to generate outcomes.

Unsupervised learning may in turn play a growing role in finding novel treatments to complex multifactorial diseases.[264] However, an unsupervised learning model also requires initial human input to inform the natural patterns it should find in a data set. Given that the human input itself may be biased, the unsupervised learning model may also demonstrate biased decision making. Furthermore, effective unsupervised learning models are acknowledged as more challenging to create, which may hinder the speed with which they can transform various areas of the healthcare space.

AI may be useful in helping clinicians overcome inherent biases towards making overly optimistic predictions. For example, clinicians often overestimate life expectancy by a factor of 5, while predictive models trained from larger amounts of data are expected to be more accurate.[265] However, the use of AI in the healthcare space can also result in the creation of another type of bias known as automation bias, whereby clinicians accept the guidance of automated systems, and by effect, do not search for confirmatory evidence themselves. A review of empirical studies of complacency and bias in human interaction with automated decision support systems showed that automation complacency and bias is more likely to occur when manual tasks compete with the automated tasks for the operator's attention. The study specifically found that operators who were responsible for many functions were more likely to reallocate attention away from automated tasks towards other manual tasks in cases of high workload, resulting in increased automation failures.[266] Given that the review did not evaluate the impact of human bias on working with AI in the healthcare space specifically, further research may be needed to determine the generalizability of such findings in the healthcare space.

Data Privacy

Society places great value on individual rights, particularly in regards to protection of personal information. In many cases, individuals' health care records may be viewed as the most

---

[262] Jiang.

[263] Deo, Rahul C. "Machine learning in medicine." *Circulation* 132, no. 20 (2015): 1920-1930.

[264] Deo. *ibid.*

[265] Avati, Anand, Kenneth Jung, Stephanie Harman, Lance Downing, Andrew Ng, and Nigam H. Shah. "Improving Palliative Care with Deep Learning." *BMC Medical Informatics and Decision Making* 18, no. 4 (2018): 122.

[266] Parasuraman, Raja, and Dietrich H. Manzey. "Complacency and Bias in Human use of Automation: An Attentional Integration." *Human Factors* 52, no. 3 (2010): 381-410.

intimate and sensitive form of personal information. Respecting an individual's right to privacy may be justified by the principle of respect for autonomy. To respect a person's privacy may be equivalent to respecting their autonomous wishes to have information about themselves released. Other ethical justifications for protecting patients' privacy includes hindering the development of potential social and psychological damages that may result from unwanted disclosures of personal health care information. Relatedly, maintaining patients' right to privacy may prevent them from suffering embarrassment and social discrimination that can result from unexpected exposure of personal information.[267]

Implementing AI in healthcare may pose challenges to maintaining patient confidentiality. A recent study by University of California, Berkeley showed that when training an AI with health information, stripping all identifying information is not enough to protect individuals' private information from being identified. While HIPAA regulations make some health care data private, it does not cover as much information as people might think, producing feasible ways for companies buying the health data to identify individuals' health information. For example, if Facebook wanted to know individual's health information, it could hypothetically buy healthcare data from a company that claimed to strip the data of all identifying information, and use step counting data tracked via the Facebook app on individuals' smartphones to figure out individuals' specific health information. In recent years, DeepMind Health, an artificial intelligence company that accesses millions of UK patients' health records to improve physicians' diagnostic ability, has merged with Google, eliciting concerns about potential threats to patients' privacy. Google has provided DeepMind Health the necessary capital to expand the AI firm's impact to many more hospitals, which has benefited significantly more hospitals, physicians, and patients globally. While DeepMind Health states that its information is encrypted, and will never be connected to Google accounts or services, many experts feel there still exist ways for Google to backtrack and pair medical records with data collected through its search engine and Gmail app.[268] The merger also allowed Google to dismantle an independent review board created to oversee Google's work with the healthcare sector, removing barriers that may have previously prevented Google from misusing private patient information.[269] Advances in companies' ability to extract private health data may in turn make it easier for certain companies to use private health information in a discriminatory basis.[270] For example, when evaluating individuals' candidacy for a job, companies may access and use individuals' private health data to unfairly make hiring decisions. Additionally, cyber criminals

---

[267] Gostin, Lawrence. "Health Care Information and the Protection of Personal Privacy: Ethical and Legal Considerations." *Annals of Internal Medicine* 127, no. 8, Part 2 (1997): 683-690.

[268] Kahn, Jeremy, and John Lauerman. "Google Taking Over Health Records Raises Patient Privacy Fears." November 12, 2018. Accessed April 11, 2019. https://www.bloomberg.com/news/articles/2018-11-21/google-taking-over-health-records-raises-patient-privacy-fears.

[269] Hern, Alex. "Google 'Betrays Patient Trust' with DeepMind Health Move." *The Guardian*. November 14, 2018. Accessed April 11, 2019. https://www.theguardian.com/technology/2018/nov/14/google-betrays-patient-trust-deepmind-healthcare-move.

[270] Hickey, John. "Artificial Intelligence Advances Threaten Privacy of Health Data." *EurekAlert!*, UC Berkeley , www.eurekalert.org/pub_releases/2019-01/uoc--aia010319.php

and security agencies can also penetrate devices that use AI models to function, yielding potentially disastrous outcomes for patients.[271]

Furthermore, given that AI systems are only increasing in the complexity of the tasks they perform, it is probable that AI systems will require multiple developers to jointly train and implement AI models. This, in turn, will likely require sharing input data sets amongst many entities, which can pose additional challenges to data protection and privacy.[272] Recent development of systems that allow multiple entities to train AI models without sharing input datasets may point to a potential solution to this conflict.[273] Another potential solution may necessitate the adoption of decentralized "federated" learning approaches, which allow data to remain local during training.[274] Developers may also use cryptonets, which are deep learning networks that are trained on encrypted data. Cryptonets can make encrypted predictions that can only be decrypted by the owner, ensuring confidentiality.[275] While the use of cryptonets still enables certain individuals to have knowledge of patients' confidential data, the goal should not be to maintain 100% confidentiality, but instead to simply work towards minimizing the number of parties who have access to the data.

Informed Consent

A central component to effective healthcare delivery is the concept of informed consent, which involves a patient agreeing on a certain treatment after completely understanding the facts and implications of related actions. To receive informed consent from a patient, a health care provider must provide all information necessary to allow the patient to choose whether to continue with a given treatment. A health care provider must specifically explain the nature of the treatment, reasonable alternatives to the proposed intervention, and the risks, benefits, and uncertainties associated with the recommended treatment.[276]

When considering AI's increased adoption in the healthcare space, questions may also emerge regarding how the dynamics of informed consent might change. The use of artificial intelligence may introduce another entity in the traditional patient-physician relationship, that can pose ethical challenges regarding how physicians should discuss the use of artificial

---

[271] Mantovani, Eugenio, Joan Antokol, Marian Hoekstra, Sjaak Nouwt, Nico Schutte, Pēteris Zilgalvis, J-P. Castro Gómez-Valadés, and Claudia Prettner. "Towards a Code of Conduct on Privacy for mHealth to Foster Trust Amongst Users of Mobile Health Applications." In *Data Protection and Privacy:(In) Visibilities and Infrastructures*, pp. 81-106. Springer, Cham, 2017

[272] Hosny, Ahmed, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz, and Hugo JWL Aerts. "Artificial Intelligence in Radiology." *Nature Reviews Cancer* (2018): 1.

[273] Shokri, Reza, and Vitaly Shmatikov. "Privacy-preserving Deep Learning." In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1310-1321. ACM, 2015.

[274] Konečný, Jakub, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. "Federated Learning: Strategies for Improving Communication Efficiency." *arXiv preprint arXiv:1610.05492* (2016).

[275] Gilad-Bachrach, Ran, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. "Cryptonets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy." In *International Conference on Machine Learning*, pp. 201-210. 2016.

[276] Schulz, Peter J., and Sara Rubinelli. "Arguing 'For' the Patient: Informed Consent and Strategic Maneuvering in Doctor–Patient Interaction." *Argumentation* 22, no. 3 (2008): 423-432.

intelligence with patients. Such a conflict may be further imagined through the following hypothetical case, published by researchers Schiff et al. in the JAMA Journal of Ethics: Mr.K is a 54-year-old man referred to Dr. L's outpatient spine neurosurgery pain in the left-sided lower back pain, left leg weakness, and shooting pain. Prior to Mr K's appointment, Dr. L reviewed the MRI of Mr. K's lumbar spine, and observed it to be a classic case that can be resolved through surgery. Dr. L communicates this recommendation to Mr. K, who responds with concerns over the surgery. To alleviate Mr. K's concerns, Dr. L responds, indicating that she will use an AI surgical tool, Mazor Robotics Guidance System, to analyze the images, and plan placement of surgical tools. Mr. K expresses discomfort over the use of a robot doing surgery, and claims that he wants Dr. L to complete the entire surgery. Dr. L does not know how to respond.

One way to resolve this dilemma may be to help make Mr. K more knowledgeable about the technology itself to increase his comfort level with it. However, given the black-box problem, which defines human uncertainty regarding how the specific mechanisms through which AI systems derives its outputs, issues may arise regarding the extent to which Mr. K can be feasibly informed about how technology works.[277] This can, in turn, hinder the extent to which it is possible for Dr. L to acquire informed consent from Mr. K. For example, how should Dr. L communicate possible biases, risks, and error rates during the informed consent process if Dr. L herself does not know all the answers due to the inherent uncertainty associated with how AI models work that may not exist with other types of technologies?

One might say that as long as Dr. L informs the patient about the error rates and minimal knowledge she has about the technology, she has successfully informed the patient. However, if Dr. L responds honestly, that she does not know exactly how the technology works, Mr. K may feel less comfortable using the technology. This question is particularly made complicated when considering that patients may already have heightened uneasiness about the use of AI in healthcare. For example, a 2016 survey of 12,000 individuals across Europe, Asia, and Africa found that 47% of people would be willing to have a robot perform minor non-invasive surgery, but only 37% would allow a robot to perform major, invasive surgeries.[278] These findings demonstrate a sizable proportion of the population might have baseline uneasiness about the complete use of medical AI to perform a task, and subsequently raises the question of the degree to which physicians should be transparent about the technology's black-box phenomenon. If Dr. L knows that ultimately, the technology has a greater efficacy rate than a human might, the question may even come up of whether Dr. L *should* communicate the uncertainty associated with how the technology works at all.

Furthermore, Dr. L may choose to communicate the use of the technology as a tool rather than the principal decision maker for the surgery. However, given that physicians can become overly reliant on the decisions made by the AI systems, perhaps such a response might not be a completely honest portrayal of the degree to which the AI will actually be the leading decision maker in a given health procedure.[279] For example, if Dr. L is likely to view any treatment recommendation made by the AI as accurate, regardless of whether it is sound, one might argue that the AI will always serve as the principal decision maker, even if it requires humans to be

---

[277] Schiff, Daniel, and Jason Borenstein. "How Should Clinicians Communicate With Patients About the Roles of Artificially Intelligent Team Members?." *AMA Journal of Ethics* 21, no. 2 (2019): 138-145.
[278] Müller, Vincent C., and Nick Bostrom. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion." In *Fundamental Issues of Artificial Intelligence*, pp. 555-572. Springer, Cham, 2016.
[279] Parasuraman.

used. This raises the question of whether Dr. L should tell Mr. K about the existence of automation bias in order to inform the patient about potential risks as thoroughly as possible. Furthermore, given that the black-box problem refers to a knowledge deficit regarding how exactly the technology works, physicians may not be able to explain how/why potential errors occur, should they happen. This raises questions of who is to blame if an error occurs. If is it not possible for the physician to know entirely how the technology works, is it right to blame him or her for potential errors? Such questions highlight the "problem of many hands" which refers to the challenges of attributing moral responsibility when the cause of a harm is distributed among many people and/or organizations in ways that obfuscates blame attribution.[280]

To minimize the creation of blame attribution conflicts, it may be necessary for all major stakeholders to take preventive steps. For example, coders and designers could implement strategies that make the technology and its underlying processes, such as how it learns from training data, explainable. Furthermore, medical device companies could clearly articulate any prerequisites for successfully applying the AI technology, such as the quality of diagnostics, and preparation for surgical procedures. Medical device companies could also make clear to health professionals and hospitals the various types of errors and side effects that may occur, and differences in predictive accuracy and error rates across demographic subgroups, health conditions, and patient histories, insofar as the information is available. Health care professionals could be responsible for acquiring basic knowledge about how the AI device works, and could be especially educated about the types and likelihood of certain errors across subgroups. Physicians should also be responsible for communicating the information to patients and health care teams. It is important to mention that many patients may not *want* to known the technicalities of how the AI system works. In cases such as these, physicians could still ask their patients through numerous methods if they have any questions or concerns about the use of AI in treatment and diagnosis. Relatedly, the possibility that a patient may not want to know how the AI system works should not absolve the physician of the responsibility to acquire basic knowledge about the technology. Finally, hospitals and health care systems could be responsible for ensuring adequate development, implementation, and monitoring of protocols and practices for the use of AI systems in healthcare. Hospitals could specifically provide training to health care providers using AI systems, which an emphasis on how physicians should inform patients about the technology.[281]

Conclusion

While there is immense potential for AI to revolutionize healthcare, it can also propagate existing medical bias, pose threats to patients' privacy rights, and complicate the process of receiving informed consent. To help ensure that the impacts of AI in healthcare are positive, targeted policies must be implemented that regulate the development and implementation of healthcare AI.

Recommendation to Limit Bias

---

[280] Thompson, Dennis F. "Moral Responsibility of Public Officials: The Problem of Many Hands." *American Political Science Review* 74, no. 4 (1980): 905-916.

[281] Schif.

To help prevent AI systems from propagating medical biases that occur due to limited representation of certain minority groups in clinical datasets, policies must be created and enforced by technology companies that require all employees developing AI intended to be used as a diagnostic or treatment tool to train AI systems using diverse data sets or use other algorithmic tools to overcome these limitations. If diverse data sets are not available, technology companies should take additional measures to ensure that they have adequate knowledge about the particular limitations of the data set and subsequently, the AI model itself. Furthermore, technology companies should facilitate collaboration between programmers and medical professionals that help programmers overcome knowledge deficits regarding ways in which the data set may be limited. For example, limited medical literature on the prevalence of a specific disease condition in certain minority groups may warrant programmers to consult with a team of healthcare professionals experienced in the specific disease, who can help tailor the development of the AI model in beneficial ways. Furthermore, when the model is complete and ready to be sold to medical device companies and hospitals, a regulatory network should require technology companies to make obvious limitations of the model and subsequently, the extent to which the device may propagate existing biases. By being transparent about the limitations of the AI model, technology companies and programmers can help create a more honest and ethical landscape for AI to make positive changes in healthcare.

Healthcare providers' may show biased decision making in their use of AI systems to make diagnosis and treatment recommendations. Specifically, healthcare professionals can be victim to automation bias, in which they become overly-reliant on decisions made by AI systems due to the models high rates of accuracy. To help healthcare providers use AI systems as tools that assist decision making instead of replace it all together, health care professionals should be trained about how AI systems are made, and subsequent limitations to the model itself. Training workshops should be required before an healthcare professional uses an AI system as a diagnostic and treatment tool. Furthermore, training should be administered and enforced by hospitals and other medical institutions that employ health care professionals.

Recommendations to Protect Privacy

Effective AI systems require access to large amounts of patient data, which can potentially violate patients' rights to privacy. To help prevent data breaches, it may be necessary for HIPPA to create stricter regulations protecting patients' health care data. Furthermore, given that current HIPPA standards require healthcare professionals to be trained on HIPPA's Privacy Rules, HIPPA should also require all developers of AI diagnostic and treatment models to be similarly trained on the rules, and enforce penalties for breaking any privacy standards. Technology companies should be required to enforce HIPPA training for such employees. While stricter regulations might make it harder to create novel AI health care tools, such standards may be justified by the weightier outcome of protecting patients' healthcare data from possible abuse.

Recommendations to Protect Standards of Informed Consent

The black-box phenomenon, in which the specific mechanisms through which AI models derive outputs are unknown to even the programmer herself, raises questions about the extent to which healthcare professionals can inform patients about the AI models used to treat and/or diagnose them. To help minimize physician and consequent patient confusion, healthcare professionals using AI systems should be required to gain sufficient knowledge about how the AI

tool works, and associated risks and limitations of the model, to explain it to an educated patient. Specifically, healthcare providers should know about the types and likelihood of errors that could potentially occur if the AI model is used. Furthermore, the health care professional should be required to ask patients through multiple methods if they have any concerns about use of the AI system itself. Hospitals and other medical institutions employing healthcare professionals should be required to host such training workshops. Training workshops should emphasize teaching healthcare professionals about various ways in which the providers can communicate about complicated technology to curious and concerned patients.

Works Cited

Anderson, Michael, and Susan Leigh Anderson. "How Should AI Be Developed, Validated, and Implemented in Patient Care?" *AMA Journal of Ethics* 21, no. 2 (2019): 125-130.

Avati, Anand, Kenneth Jung, Stephanie Harman, Lance Downing, Andrew Ng, and Nigam H. Shah. "Improving Palliative Care with Deep Learning." *BMC Medical Informatics and Decision Making* 18, no. 4 (2018): 122.

Challen, Robert, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atanasova. "Artificial Intelligence, Bias and Clinical Safety." *BMJ Qual Saf* 28, no. 3 (2019): 231-237.

Chen, Irene Y., Peter Szolovits, and Marzyeh Ghassemi. "Can AI Help Reduce Disparities in General Medical and Mental Health Care?." *AMA Journal of Ethics* 21, no. 2 (2019): 167-179.

Coakley, Meghan, Emmanuel Olutayo Fadiran, L. Jo Parrish, Rachel A. Griffith, Eleanor Weiss, and Christine Carter. "Dialogues on Diversifying Clinical Trials: Successful Strategies for Engaging Women and Minorities in Clinical Trials." *Journal of Women's Health* 21, no. 7 (2012): 713-716.

Curry, Brian. "An Introduction to Transfer Learning in Machine Learning." *Medium*. July 26, 2018. Accessed April 11, 2019. https://medium.com/kansas-city-machine-learning-artificial-intelligen/an-introduction-to-transfer-learning-in-machine-learning-7efd104b6026.

Deo, Rahul C. "Machine Learning in Medicine." *Circulation* 132, no. 20 (2015): 1920-1930.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks." *Nature* 542, no. 7639 (2017): 115.

Gershgorn, Dave. "If AI Is Going to Be the World's Doctor, It Needs Better Textbooks." Prescription AI. September 10, 2018. Accessed April 21, 2019. https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks/.

Gilad-Bachrach, Ran, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. "Cryptonets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy." In *International Conference on Machine Learning*, pp. 201-210. 2016.

Gostin, Lawrence. "Health Care Information and the Protection of Personal Privacy: Ethical and Legal Considerations." *Annals of Internal Medicine* 127, no. 8, Part 2 (1997): 683-690.

Hern, Alex. "Google 'Betrays Patient Trust' with DeepMind Health Move." *The Guardian*. November 14, 2018. Accessed April 11, 2019. https://www.theguardian.com/technology/2018/nov/14/google-betrays-patient-trust-deepmind-healthcare-move.

Hickey , John. “Artificial Intelligence Advances Threaten Privacy of Health Data.” *EurekAlert!*, UC Berkeley , www.eurekalert.org/pub_releases/2019-01/uoc--aia010319.php.

Hosny, Ahmed, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz, and Hugo JWL Aerts. "Artificial Intelligence in Radiology." *Nature Reviews Cancer* (2018): 1.

Huang, T. "Computer Vision: Evolution and Promise" from CERN Document Server (1996).

Jiang, Fei, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. "Artificial Intelligence in Healthcare: Past, Present and Future." *Stroke and Vascular Neurology* 2, no. 4 (2017): 230-243.

Kahn, Jeremy, and John Lauerman. "Google Taking Over Health Records Raises Patient Privacy Fears." November 12, 2018. Accessed April 11, 2019. https://www.bloomberg.com/news/articles/2018-11-21/google-taking-over-health-records-raises-patient-privacy-fears.

Koch, Marta. "Artificial Intelligence is Becoming Natural." *Cell* 173, no. 3 (2018)

Konečný, Jakub, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. "Federated Learning: Strategies for Improving Communication Efficiency." *arXiv preprint arXiv:1610.05492* (2016).

Mantovani, Eugenio, Joan Antokol, Marian Hoekstra, Sjaak Nouwt, Nico Schutte, Pēteris Zilgalvis, J-P. Castro Gómez-Valadés, and Claudia Prettner. "Towards a Code of Conduct on Privacy for mHealth to Foster Trust Amongst Users of Mobile Health Applications." In *Data Protection and Privacy:(In) Visibilities and Infrastructures*, pp. 81-106. Springer, Cham, 2017

Müller, Vincent C., and Nick Bostrom. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion." In *Fundamental Issues of Artificial Intelligence*, pp. 555-572. Springer, Cham, 2016.

Murthy, Vivek H., Harlan M. Krumholz, and Cary P. Gross. "Participation in Cancer Clinical Trials: Race-, Sex-, and Age-based Disparities." *JAMA* 291, no. 22 (2004): 2720-2726.

Parasuraman, Raja, and Dietrich H. Manzey. "Complacency and Bias in Human Use of Automation: An Attentional Integration." *Human Factors* 52, no. 3 (2010): 381-410.

Schiff, Daniel, and Jason Borenstein. "How Should Clinicians Communicate With Patients

About the Roles of Artificially Intelligent Team Members?." *AMA Journal of Ethics* 21, no. 2 (2019): 138-145.

Schulz, Peter J., and Sara Rubinelli. "Arguing 'For' the Patient: Informed Consent and Strategic Maneuvering in Doctor–Patient Interaction." *Argumentation* 22, no. 3 (2008): 423-432.

Shokri, Reza, and Vitaly Shmatikov. "Privacy-preserving Deep Learning." In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1310-1321. ACM, 2015.

Somashekhar, S. P., R. Kumarc, A. Rauthan, K. R. Arun, P. Patil, and Y. E. Ramya. "Abstract S6-07: Double Blinded Validation Study to Assess Performance of IBM Artificial Intelligence Platform, Watson for Oncology in Comparison with Manipal Multidisciplinary Tumour Board–First Study of 638 Breast Cancer Cases." (2017): S6-07.

Soni, Devin. "Supervised vs. Unsupervised Learning." *Towards Data Science*. March 22, 2018. Accessed April 11, 2019. https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d.

Steadman, Ian. "IBM's Watson Is Better at Diagnosing Cancer than Human Doctors." WIRED. October 04, 2017. Accessed April 20, 2019. https://www.wired.co.uk/article/ibm-watson-medical-doctor.

Thompson, Dennis F. "Moral Responsibility of Public Officials: The Problem of Many Hands." *American Political Science Review* 74, no. 4 (1980): 905-916.

US FDA, September. "Guidance for Industry, Electronic Source Data in Clinical Investigations." (2017).

Is Virtual Reality a Viable Form of Treatment for Anxiety Disorders?

Sheridan Wilbur

[Stacy[282]], a 26 year old African American female, was an executive for a large financial institution in NYC and was presented for a psychological evaluation approximately four months after the attack on the World Trade Center.[283] She worked in a competitive industry, but thrived under pressure and was "on the fast track" for success. Stacy described herself as having bright prospects and high ambition but on September 11, 2001, all of that changed. She was across the street from the North tower when the first plane hit the building. Her family described her as distant and cut off after the 9/11 attacks and it was her mother who initially called and asked for help, noting that her daughter was "not herself and unusually irritable."

Stacy underwent a psychological assessment and spoke in a monotone voice. She described her experience with little emotion and denied feeling any terror. She was diagnosed with PTSD and co-morbid major depression, and met the criteria for moderate to severe symptoms in each of the DSM-IV cluster areas for PTSD. Stacy had re-experiencing symptoms such as intrusive imagery of the plane striking the tower, the building collapsing and felt distressed when she was confronted with reminders. Her avoidant symptoms included avoiding thoughts of the attack, refusing to watch TV or read newspapers and staying away from situations that she perceived as vulnerable (refusing to stay at her boyfriends high rise apartment).

Stacy began working with a therapist to receive PTSD treatment and underwent traditional imaginal exposure based therapy (EBT). She was asked to mentally revisit and imagine her experience from 9/11 and recall the events to her therapist in detail and in present tense. They were hopeful that Stacy would repeatedly describe her experience until her feelings of fear diminished, but she failed to make progress after four sessions. She continued to describe her trauma with a flat, emotionless affect and scored a zero on scales that reflected her emotional engagement with the trauma. Stacy was frustrated with this treatment plan, despite telling her therapist that she was 'fine' and recognized something was wrong because she still felt distressed by her irritability. It is common for people with PTSD to shut down emotionally or struggle to visualize their memory and connect with it because of its' traumatic nature.

However, Stacy's therapist suggested a less traditional form of exposure therapy that would not require her to imagine these memories, and offered virtual reality (VR) exposure therapy instead. Virtual reality exposure therapy (VRET) was an appealing alternative for Stacy because it does not require her to imagine the trauma on her own and offers the potential to facilitate emotional engagement by using additional sensory inputs. Trauma gets stuck in the non-analytical parts of the brain — the part that holds emotions, creativity, experiences, art, etc but trauma is image based, somatic and nonverbal. In order for a patients with PTSD to heal, they need to access the left hemisphere of the brain— the part that deals with logic, reasoning and language and put an analytical narrative to their trauma.[284]

---

[282] *Name changed for privacy.

[283] Difede, Joann, and Hunter G. Hoffman. "Virtual Reality Exposure Therapy for World Trade Center Post-Traumatic Stress Disorder: A Case Report." *Cyberpsychology & Behavior* 5, no. 6 (2002): 529-535.

[284] Stinnett, Randy. "How to Treat Emotional Trauma." *Psychiatry & Psychotherapy Podcast*.

Therefore, the additional sensory and 3D image inputs from VRET could facilitate patients like Suzy to have access the emotional parts of their experience that might have been stuck from visualization. VR's additional imagery could help patients put words to their trauma and offer the possibility to improve the efficacy of exposure therapy for PTSD by helping to integrate both parts of the brain.

Stacy agreed to try VRET and wore a VR helmet that placed two miniature computer screens in front of her eyes. Position tracking devices signaled any changes in Stacy's movement to the computer and the scenery in VR changed as she moved her head orientation (i.e. virtual objects would get closer as Stacy leaned forward in the real world). This way, Stacy didn't rely on her imagination to retell the 9/11 attacks and VR placed her into a computer generated environment that simulated her experience in lower Manhattan. She put on the head gear and began treatment by viewing the Twin Towers from a distance and with no sound effects. Stacy was administered six sequences that gradually became more complete and closer to her experience on 9/11, such as adding sound effects, a jet flying over, crashing with explosion, another jet flying over, collapsing tower, people jumping, until it was the closest, most realistic sequence that replicated her specific trauma. Stacy's therapist controlled what she experienced by using keys on the keyboard and was able to simultaneously view what Stacy saw on a computer screen nearby.

She spent about 45 to 60 minutes in VR per session, and dictated the pace on her own. Stacy was required to repeat each sequence until habituation occured and her 'Subjective Units of Distress' level decreased by at least 50%, meaning her anxiety was reduced and she was naturalized to witnessing the events in the sequence without the same degree of fear. This approach was designed to evoke discomfort but not to the point that it was intolerable and Stacy provided informed consent before she approached the next sequence. Despite her inability to emotionally engage in a narrative using imaginal exposure before, it was evident that she engaged immediately with the virtual WTC world by measures in her verbal report, behavior and physiological signs of emotional arousal.

Stacy started VR treatment with nervousness, but according to the therapist, she had "determined anticipation."[285] She viewed the Twin Towers through the headset, and began to cry for the first time, and said that "she'd never thought she'd be able to look at them again." The image provoked Stacy to recall her experience in an emotional way and share memories that were previously inaccessible. She described the harrowing day in detail to her therapist; from the moment that she had tried to run but was held back by a crush of bodies who were trying to escape, to the heartbreaking decision she made after freeing herself from beneath other people. Stacy recalled the choice that she was required make when debris was falling all around and was unable to help a woman who cried out for help. She remembered that she had no shoes on, no money and her feet were bleeding. Her traumatic experience did not end until she met her boyfriend in midtown Manhattan and was taken home.

[285]Difede, Joann, and Hunter G. Hoffman. "Virtual Reality Exposure Therapy for World Trade Center Post-Traumatic Stress Disorder: A Case Report." *Cyberpsychology & Behavior* 5, no. 6 (2002): 529-535.

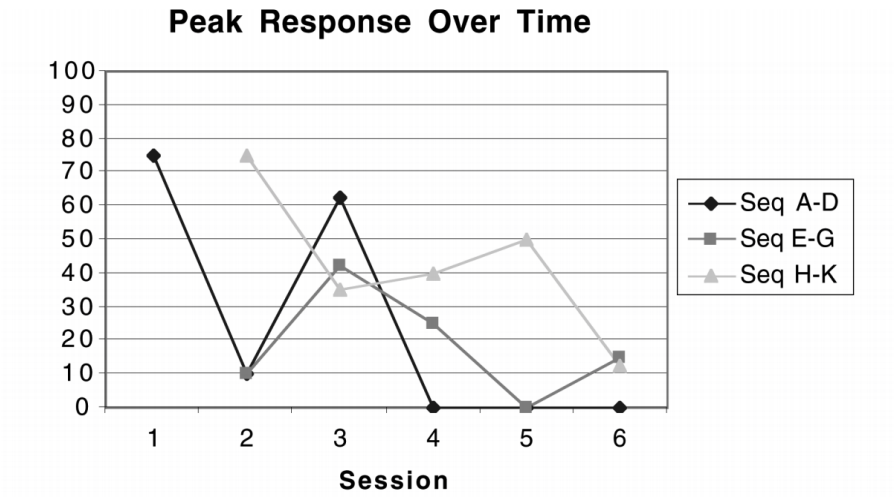

FIG. 1. Event 1: (a) A jet flies over the WTC towers, but doesn't crash, normal NY city street sounds, (b) Then a jet flies over, hits building, but no explosion, (c) Then a jet flies over, crashes with explosion but no sound effects, and (d) Then a jet flies over, crashes with explosion and explosion sound effects. Event 2: (e) Burning smoking building (with hole where jet crashed), no screaming, (f) Burning smoking building (with hole where jet crashed) and screaming, and (g) Burning smoking building (with hole where jet crashed), screaming and people jumping. Event 3: (h) Second jet crashes into second tower with explosion and sound effects, (i) Second tower collapses with dust cloud, (j) First tower collapses with dust cloud, and (k) The full sequence.

\

Image from: Difede, Joann, and Hunter G. Hoffman. "Virtual Reality Exposure Therapy for World Trade Center Post-Traumatic Stress Disorder: A Case Report." *Cyberpsychology & Behavior* 5, no. 6 (2002): 529-535.

Stacy underwent six VR exposure sessions to prime her to access those traumatic moments on 9/11, and after fourteen weeks of treatment, she was able to remember what happened in greater detail. Over time, Stacy's peak response when entering the virtual world gradually improved (see graph) and she was able to emotionally engage in exposure based therapy, rather than deny feelings of fear and vividly recall her experience. She did not have the same amount of detachment or underlying terror when retelling the events to her therapist and it was clear from examination by an independent assessor, verbal reports, and a self report measure using the Subjective Units of Distress scale, that Stacy no longer met criteria for PTSD, Major Depression or any other psychiatric disorder. She claimed that it was VR that helped her access her memories and manage her symptoms of anxiety.

Introduction

Stacy is one particular case study, but her experience is a compelling instance that demonstrates VRET technology as a promising alternative to standard, human driven exposure based therapy for PTSD. There are additional studies like Stacy's that also demonstrate the effectiveness of VR for PTSD and other anxiety disorders that I will highlight below. In the following chapter, I investigate immersive VR exposure therapy in more detail and explore if this form of technology could be used to facilitate emotional engagement, overcome limitations from real world and imaginal exposure, and provide more control or customization for clinicians and patients. VRET is an alternative form of therapy that holds the possibility to potentially improve the efficacy of exposure therapy for anxiety disorders, but I acknowledge the limitations of my claims and consider the ethics of why and if, it should be used.

First, I define virtual reality and anxiety disorders and describe the standard forms of psychological treatment. Next, I provide more details on exposure based therapy (EBT) and include evidence towards why it's an efficacious treatment for anxiety disorders. After, I specifically describe the standard forms of EBT prior to VR; in vivo (real world) and imaginal. I outline their limitations, and mainly, argue that exposure treatment is limited by the patient's degree of emotional engagement. I propose that VR exposure therapy is an ethical and efficacious way to increase emotional engagement that is inadequate in in vivo or imaginal exposure, and suggest how VR can amplify patients degree of presence, control and customization. I define VR in detail, explain how VRET works and offer examples of VR therapy on the market for anxiety disorders today. The last section of this chapter investigates the ethical issues of virtual reality as a form of therapy, considers what technology can add to treatment that standard treatment cannot, and consider what effects that VR has on the future of patients, clinicians and the burden of global mental illness.

What is Virtual Reality (VR)?

Virtual reality is a technological interface that allows users to experience computer-generated environments within a controlled setting.[286] VR uses the capability of computers to synthetically introduce stimuli to the user's senses by creating a 3D graphical environment from quantitative data. This technology includes input devices that sense the subject's reactions and motions. The computer can "modify the synthetic environment accordingly and creates the illusion of interacting with, and thus being immersed within the environment."[287] As a result, users are able to interact and navigate through the imaginary world in front of them and change their perception on a more 'complete' and immersive level by engaging in simulations for flying, walking, socializing, gaming, public speaking, phobias, etc. This technology is made possible by using a head-mounted display or a hand-held controller to track head movements so the movements and images change in a natural way, allowing for the greatest sense of presence or immersion. It's important to note that with VR, it is the user's perception that is changing and not their reality.

History of VR

Early development of VR began in the 1950s and 60s as technology to engage the user's senses. In 1957, the Sensorama Machine was released and later patented as an interactive film experience in which viewers were invited to watch a film that would use all their senses. In 1968, Ivan Sutherland developed the first head-mounted display (HMD), and users were able to view 3D objects in a more intimate way, and have a personal experience with VR. By the early 1980s, Myron Krueger developed the first program that allowed people to interact and change computer generated images through bodily movements, and coined the term 'Artificial Reality' to describe "the ultimate expression of this concept." Krueger was the first person who suggested VR was a viable treatment for mental health disorders, but Jaron Lanier was the first to use the name

[286] Maples-Keller, J. L., Bunnell, B. E., Kim, S. J., & Rothbaum, B. O. (2017). The Use of Virtual Reality Technology in the Treatment of Anxiety and Other Psychiatric Disorders. *Harvard review of psychiatry*, *25*(3), 103–113. doi:10.1097/HRP.0000000000000138

[287] Gorini, Alessandra, and Giuseppe Riva. "Virtual Reality in Anxiety Disorders: the Past and the Future." Expert Review of Neurotherapeutics, vol. 8, no. 2, 2008, pp. 215–233., doi:10.1586/14737175.8.2.215.

'virtual reality.' Lanier was later the founder of the visual programming lab (VPL), a company that developed a range of virtual reality gear such as the Dataglove, the EyePhone head mounted display and Virtual Reality goggles.

Today, Virtual Reality has advanced to become more lightweight and practical, and is sold at a more affordable price (Oculus Go, $199; Google Cardboard, $15). These advancements possibly offer a technological solution for anxiety that is more accessible and effective than standard psychological treatment alone. VR is a popular technology used along with Augmented Reality (AR) and Artificial Intelligence (AI), but there remain limited studies on *intelligent* virtual environments for anxiety disorders. Intelligent virtual environments would apply AI to VR by creating environments that provide knowledge "to direct or assist the user rather than relying entirely on the user's knowledge and skills, those in which the user is represented by a partially autonomous avatar, or those containing intelligent agents separate from the user."[288] Artificial Intelligence's (AI) computational power may allow VR systems to be far more responsive and usable on smaller devices, but intelligent virtual environments will need to be researched more in the future to support these claims.

Virtual Reality Exposure Therapy (VRET)

More recently, virtual reality come into the conversation as an efficacious treatment solution for PTSD and other anxiety disorders. This technology has been increasingly used in the context of mental health treatment, and within clinical research.[289] Over 20 years ago, Barbara Rothbaum, PhD. of Emory University School of Medicine, and colleagues first proved that virtual reality-based exposure therapy could help people overcome a fear of heights, and this was the first instance that VR therapy was found to be effective.[290] VRET works by exposing patients to anxiety triggers within a controlled environment but is distinct from in vivo or imaginal exposure because it remains as a simulated reality, and only changes the patient's perception of reality while undergoing treatment. Jeremy Bailenson, founding director of the Virtual Human Interactive Lab at Stanford University argues that "our lab studies have shown that, in general, the brain tends to treat a VR event in a similar way to an actual event. VR, unlike a normal video game, activates the motor cortex and the perceptual system in a similar way to real life experiences. So there's a fundamental difference between performing an event in a traditional video game or watching a movie and doing it inside VR."

VR utilizes experiential learning to give people an active, not passive, opportunity to explore a more threatening space and *safely* overcome their fear. However, similar to other forms of exposure therapy, the aim of VR is still to condition patients to respond positively to events that bring about their particular anxiety, rather than experiencing physiological stress associated

---

[288] Luck, M., & Aylett, R. (2000). Applying artificial intelligence to virtual reality: Intelligent virtual environments. *Applied Artificial Intelligence, 14*(1), 3-32. doi:10.1080/088395100117142

[289] Maples-Keller, J. L., Bunnell, B. E., Kim, S. J., & Rothbaum, B. O. (2017). The Use of Virtual Reality Technology in the Treatment of Anxiety and Other Psychiatric Disorders. Harvard review of psychiatry, 25(3), 103–113. doi:10.1097/HRP.0000000000000138.

[290] Rothbaum, Barbara, et al. "Effectiveness of Computer-Generated (Virtual Reality) Graded Exposure in the Treatment of Acrophobia." American Journal of Psychiatry, vol. 152, no. 4, 1995, pp. 626–628., doi:10.1176/ajp.152.4.626.

with the original trauma that caused their anxiety disorder. As a result, limitations of EBT as an efficacious solution should not be overlooked, such as treatment adherence.[291]

What is Anxiety?

Anxiety can best be described as the butterflies in your stomach before a job interview, the tension you feel during a confrontation, or the way your heart pounds when you are in danger. It's the sweating you might feel before a big test or presentation or the heat you feel on your face after an awkward social situation, but this feeling is normal. Anxiety is a universal feeling of distress, and is often a helpful response that prepares you for action, but this 'fight-or-flight' feeling becomes Anxiety Disorder when it creates persistent, excessive and intrusive worrying and interferes with daily functioning (DSM-5).

Anxiety Disorder (AD)

There are five major types of anxiety disorders listed in the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition (DSM-5), PTSD being one of them. The manual also includes Generalized Anxiety Disorder, Social Anxiety Disorder, Separation Anxiety Disorder, and Panic Anxiety Disorder as forms of AD. These forms of mental illness are not rare and over 40 million adults in the United States are affected by Anxiety Disorder every year. Patients report considerable fear that prevents them from doing a normal activities, such as driving, taking a plane, interacting with strangers, or staying in a crowded place. They report constant, unsubstantiated worry that leads them to avoid social situations for fear of being judged, embarrassed or humiliated and have irrational fears that pose little realistic danger.[292] They also might have recurring flashbacks related to a traumatic event that occurred years before. Despite various forms of treatment, only about ¼ of patients seek help, perhaps due to stigma and cost and there is no evidence that the prevalence rates have changed recently.[293] As a result, Anxiety Disorder remains the most common form of mental illness in the United States. If "what you think, you become" is true, anxiety disorder has debilitating implications on patients who are suffering, decreasing their ability to engage in daily activities, increasing their risk of physical illness and influencing how they interact in the social world.

Treatment for Anxiety Disorder

Fortunately, Anxiety Disorder is highly treatable, and for the patients that do seek help, most respond well to treatment and return to fulfilling and productive lives. Standard psychological treatment includes medication and cognitive-behavioral therapies (CBT). CBT involves two main components: cognitive therapy and behavior therapy and focuses on identifying and understanding how negative thoughts and abnormal behavior can contribute to anxiety. CBT promotes patients to change their biased and dysfunctional beliefs by "discussing,

---

[291] Gonçalves, R., Pedrozo, A. L., Coutinho, E. S., Figueira, I., & Ventura, P. (2012). Efficacy of virtual reality exposure therapy in the treatment of PTSD: a systematic review. *PloS one*, *7*(12), e48469. doi:10.1371/journal.pone.0048469

[292] "Understand the Facts." Anxiety and Depression Association of America, ADAA, adaa.org/understanding-anxiety.

[293] Bandelow, B., & Michaelis, S. (2015). Epidemiology of anxiety disorders in the 21st century. Dialogues in clinical neuroscience, 17(3), 327-35.

testing, and modifying them." Therapists work with patients to help them understand that it is not the present situation that they are in that determines how they feel, but their perception about the situation. CBT is the most widely used therapy for anxiety disorders and there is considerable research to support it as efficacious. In over more than 15 randomized control trials, CBT was the most effective treatment compared to a placebo, wait list or active control treatment and was stable up to several years.[294]

In Stacy's case, she utlizied CBT by facing her trauma and 'testing' the situation on 9/11 through the VR headset, 'discussing' the events with her therapist after each sequence, and 'modifying' her narrative; until she formed an affective tale with appropriate emotional engagement and was able to grasp a realistic perspective on the trauma. CBT and specifically, VR-exposure therapy, facilitated Stacy's perception about the events from 9/11 to allow her to return to engaging fully with her friends, family and coworkers and cease experiencing PTSD symptoms.

Exposure Based Therapy (EBT)

Exposure based therapy is one type of CBT and repeatedly exposes patients to their fear, phobias or trauma using in vivo, imaginal and virtual reality exposure. The goal of exposure based therapy is to activate and modify fear structures by exposing a person to their specific feared situation or objects that trigger anxiety, dissociate faulty meanings to stimuli and learn to attach new, more realistic beliefs about the fear and weaken their previous negative associations. Most therapists use a graded approach and expose patients to a mildly feared stimuli first and then, over time, include more strongly feared stimuli. However, an alternative exposure based therapy is 'flooding,' where the therapist addresses the most feared stimuli first. Both approaches are equally effective but patients usually prefer the comfort that comes from a graded approach.

Patients gradually gain an increased sense of control over the situation and eventually, diminish anxiety. They learn to associate social interactions as positive and rewarding through repeated exposure, through in vivo (real life) or imaginal (requiring patients imagination) treatment. For example, patients with social anxiety disorder tend to avoid social interactions, such as public speaking or eating with others, due to a fear memory that deems them as a negative experience. However, with exposure based therapy, patients learn to give a public speech or eat a meal in a group setting and discover how much their feared expectation is actually likely to occur and how to cope when it does happen. As a result, they learn to adjust to formally distressing factors, such as sweaty palms or anxious thoughts during the experience through exposure based therapy and use this learning to apply towards real life situations.[295]

Several studies demonstrate that EBT is an efficacious treatment for anxiety disorders and in a single-session in vivo exposure (lasting 1-3 hours), patients with specific phobias had a 90% reduction in fear, avoidance, and overall level of impairment when evaluated in a follow up. After an average of 4 years, 65% of patients no longer had a phobia.[296] Roy MJ et al, researchers

---

[294] Schuurmans, Josien, et al. "A Randomized, Controlled Trial of the Effectiveness of Cognitive-Behavioral Therapy and Sertraline versus a Waitlist Control Group for Anxiety Disorders in Older Adults." The American Journal of Geriatric Psychiatry : Official Journal of the American Association for Geriatric Psychiatry, U.S. National Library of Medicine, Mar. 2006.

[295] Moulds ML, Nixon RD. In vivo flooding for anxiety disorders: proposing its utility in the treatment posttraumatic stress disorder. J Anxiety Disord. 2006;20:498-509.

[296] Öst LG. One-session treatment for specific phobias. Behav Res Ther. 1989;27:1-7.

from Uniformed Services University of the Health Sciences in Bethesda, Maryland, used VRET on soldiers diagnosed with PTSD, and claimed "PTSD is associated with alterations in regional brain function" and "overall, our findings imply that exposure therapy promotes recovery, or return to baseline, in brain regions that are associated with emotion regulation and management of stress."[297]

Limitations to In Vivo and Imaginal Exposure Therapy

However, about 40-60% of patients don't respond as well as expected from in vivo and imaginal based EBT. Researchers hypothesized this discrepancy could be due to "premature, noncompliance with exposure work within and between sessions" and could be not willing to practice CBT due to its demanding and stress-evoking nature.[298] However, the unique capabilities of VR open the possibility to include it as an efficacious form of exposure based treatment for patients with PTSD and other anxiety disorders and overcome drawbacks from traditional EBT.

There is a summary of findings to bolster this claim, such as in Difede et al, five patients did not respond to traditional exposure, but had a positive response to VRET and reduced at least 25-50% of their symptoms (according to CAPS (Clinician-Administered PTSD Scale)).[299] Wood et al examined the effect of VRET on twelve participants and found a significant reduction in their PTSD scores (according to the Post Traumatic Stress Disorder Checklist) and no longer fulfilled the criteria for PTSD.[300] In addition, a study by McLay et al exposed 20 active duty military personnel in a randomized trial and assigned ten patients to VRET and 10 patients to 'treatment as usual' (included CBT and medication). VRET protocol included psychoeducation, relaxation, attentional and autonomic control training and exposure to a VR simulation of Iraq or Afghanistan. At the end of the sessions, seven out of the ten participants who used VRET showed improvements greater than 30% on the CAPS scale while only one of the participants who underwent treatment as usual showed greater than 30% improvement.[301]

How could VR possibly improve traditional forms of Exposure Based Therapy (EBT)?

1. Increase emotional engagement
2. Without real-world risks (in vivo)
3. Increase imaginal capacities (imaginal)
4. Increase control and customization (for patient and therapist)

---

[297] Roy MJ, Francis J, Friedlander J, Banks-Williams L, Lande RG, et al. (2010) Improvement in cerebral function with treatment of posttraumatic stress disorder. Ann NY Acad Sci 1208: 142–149.

[298] Kathmann, Norbert. "Obsessive-compulsive disorder across the life span." (2015): 119-126.

[299] Difede J, Cukor J, Jayasinghe N, Patt I, Jedel S, et al. (2007) Virtual reality exposure therapy for the treatment of posttraumatic stress disorder following September 11, 2001. J Clin Psychiatry68(11): 1639–1647.

[300] Wood D, Murphy J, McLay R, Koffman R, Spira J, et al. (2009) Cost effectiveness of virtual reality graded exposure therapy with physiological monitoring for the treatment of combat related posttraumatic stress disorder. Studies in Health Technology & Informatics 144(7): 223–229.

[301] McLay RN, Wood DP, Webb-Murphy JA, Spira JL, Wiederhold MD, et al. (2011) A Randomized, Controlled Trial of Virtual Reality-Graded Exposure Therapy for Post- Traumatic Stress Disorder in Active Duty Service Members with Combat-Related Post- Traumatic Stress Disorder. Cyberpsychology, Behavior, and Social Network 14(4): 223–229.

First, VR could amplify the degree of presence a user feels by submerging themselves into a virtual reality and therefore, increase emotional engagement. Secondly, VR is without the real world risks or demands from in vivo, and circumvents the lack of imaginative capacities from imaginal exposure by using a simulated reality. Lastly, VR could increase the user's sense of control and autonomy, and improve customization for treatment that would also improve the efficacy of exposure based therapy.

I suggest these four reasons in more detail to consider if VR should be included as an alternative method for EBT and improve the recovery rates, but it is important to add a disclaimer. There are limitations to all of these arguments and we can not assume that VR will adequately treat patients who were unable to achieve success from standard EBT. Rather, we must also consider why patients are under engaging emotionally, what is keeping them averse to real world risks or self imagining, if they are attending all of their appointments and/or actively consenting to treatment. However, these questions do not undermine VRET as a possible form of EBT and further research should be conducted to bolster the argument that these four factors can indeed overcome standard EBT treatment.

The role of emotional engagement in EBT

Effective exposure therapy requires patients with PTSD, or other anxiety disorders, to repeatedly retell and emotionally engage with their trauma in the present tense. Patients successfully recover when they process their memories without the same degree of associated terror. One possible argument for the importance of emotional engagement in EBT recovery is from a study by Jaycox et al., which concluded that patients with high emotional engagement improved more than the others. The study argues that EBT is limited by the patient's ability and degree of, emotional engagement with the fear stimuli, but this study is not enough evidence for a definitive conclusion.[302] Some patients may refuse to engage in in vivo treatment because they cannot overcome their own hesitations to exposure to an actual phobic stimulus and this could explain the discrepancy in efficacious results from traditional EBT. It is a well known and defining feature in anxiety disorders to avoid things which are fearful and this could lead some patients to refuse to engage, and/or fail to improve from in vivo or imaginal exposure.

1. Virtual Reality as an aid in Emotional Engagement?

However, if emotional engagement is as important in recovery that Jaycox suggests, then VR shows promise to overcome the limitations of emotional engagement from in vivo and imaginal exposure for PTSD and could emerge as an efficacious alternative because of the degree it simulates present reality with additional sensory inputs. The sense of presence or "being there" in VR is defined as "the psychological perception of being in, or existing in the virtual environment in which one is immersed." The more one becomes present, or immersed, the more one is thought to become emotionally aroused, and can develop a stronger sense of realism that can be applied to real world scenarios when fear stimuli is provoked.[303] Presence, or immersion, in VR systems is facilitated through head mounted displays, gesture-sensing gloves, synthesized sounds and vibrotactile platforms and are used to activate multiple senses. These

[302] Jaycox, L.H., Foa, E.B., & Morral, A.R. (1998). Influence of emotional engagement and habituation on exposure therapy for PTSD. Journal of Consulting and Clinical Psychology 66:186–192.
[303] Burmester, Alex. "How Do Our Brains Reconstruct the Visual World?" The Conversation, 19 Dec. 2018.

inputs encourage active exploration of the virtual environment, thereby, aid in emotional engagement.[304]

But, to what extent do users feel as though they are "really there"? Presence, or emotional engagement is viewed as critical to having participants respond in VR as they would in reality (in vivo exposure), but is hard to measure this concept objectively or determine how different the degree of 'presence' is between watching a video, sticking on a headset or using VR goggles. Kober and Neurper attempted to quantitatively measure the user's subjective experience of presence and hypothesized that it is marked by increased attention towards relevant stimuli in the virtual environment and less attention to VR irrelevant stimuli. The researchers identified distinct ERP patterns that are associated with increased presence and corresponded to less attention towards VR irrelevant stimuli. They went on to use an EEG to measure brain patterns of the user and found stronger activation of frontal and parietal brain regions in the more immersive system. Therefore, there is preliminary scientific evidence to suggest there are differences in the degrees of presence elicited by a desktop VR system and immersive single wall VR system; and implies that there is a range of 'presence' for participants who undergo VRET, depending on the device they use.

Furthermore, in a study by Kim, Rosenthal, Zielinski and Brady, the researchers investigated the effects of different virtual environment technologies on emotional arousal and performance that could bolster our claim for emotional engagements association in DEBT recovery[305]. They found different types of VR systems result in different emotional responses and the more immersive and present systems (DiVE and HMD) yielded a higher likelihood of treatment efficacy. Immersive virtual environments where a participant stands or walks inside a room with 6 walled-projected computer displays were more effective than a simple desktop computer with a standard monitor. This study supports the idea that the type of VR technology used can influence the degree to which a user feels present or immersed, and thereby improve treatment efficacy. However, if not all VRET treatments elicit the same amount of 'engagement' or 'presence' and VR's efficacy for anxiety treatment is associated with the degree of engagement that a user experiences while in a session (evaluated by self report measures), then this study holds implications for the VR platform of choice when employing an EBT plan for anxiety disorders[306].

Current VR on the Market

In the last couple of years, VRET has exploded on the market thanks to systems that operate on cell phone apps and the proliferation of AI. Users can download an free app within minutes that incorporates EBT to possibly help manage distressing thoughts, and there are also fully immersive VR options available to treat severe anxiety disorders (PTSD etc) for those who can afford it.

---

[304] Maples-Keller, J. L., Bunnell, B. E., Kim, S. J., & Rothbaum, B. O. (2017). The Use of Virtual Reality Technology in the Treatment of Anxiety and Other Psychiatric Disorders. Harvard review of psychiatry, 25(3), 103–113.

[305] Kim, Kwanguk, et al. "Effects of Virtual Environment Platforms on Emotional Responses." Computer Methods and Programs in Biomedicine, vol. 113, no. 3, 2014, pp. 882–893., doi:10.1016/j.cmpb.2013.12.024.

[306] Juan, M. Carmen, and Dennis Joele. "A Comparative Study of the Sense of Presence and Anxiety in an Invisible Marker versus a Marker Augmented Reality System for the Treatment of Phobia towards Small Animals." International Journal of Human-Computer Studies, vol. 69, no. 6, 2011, pp. 440–453.

Currently, cell phone apps that incorporate VRET for anxiety disorders include 'Arachnophobia,' 'Richie's Plank Experience, 'Limelight' (fear of public speaking). Arachnophobia is a self guided exposure therapy app that exposes patients with an irrational fear of spiders to an increasing number of spiders in the room, and over time, condition them to being comfortable with their presence, without a "fight or flight" response kicking in. 'Richie's Plank Experience' simulates an environment where users are on a plank, 80 stories above ground, to encourage users to become less anxious about heights. 'Limelight' addresses people's fear of public speaking, and gives users the option to address an audience in a business meeting, classroom or large hall. There are also a plethora of apps that use VRET technology to help users relax by entering calming visual and sound experiences at places like the beach, forest, mountains etc.

More immersive VR technology is on the market too, and triggers other senses (i.e., sound, smell, touch) by tactile sensations, wind, temperature extreme, acoustic effects, tilting/dropping or moisture/rain using haptic devices. For example, Bravemind, a VR exposure therapy prototype developed by USC's Institute for Creative Technologies, submerges war veterans suffering from PTSD in an 'authentic 360-degree battlefield' by simulating smell and heat of artillery fire and the sound of explosions. This technology and therapy simulates the most real, yet highly controlled setting to expose patients to their fear, and aims to decrease their physiological stress response, increase their social functioning and improve their overall quality of life. So if deeper emotional engagement through technology like Bravemind does indeed facilitate recovery from EBT, fully immersive VR options should be considered as additional forms of treatment.

Another example of immersive technology for anxiety disorders is DEEP VR, a video game that uses a headset and monitors to take users into a beautiful and serene world that responds to their stress with soothing resonance[307]. DEEP VR guides users to take deep breaths, and the deeper the breaths, the further they are submerged into the landscape. DEEP VR targets patients with Anxiety Disorders such as Generalized Anxiety Disorder and Panic Disorder, incorporating EBT by encouraging users to practice tolerating, managing and eventually, overcome, feelings of anxiety. This process occurs organically, as players follow their curiosity and explore the underworld playground, as well as their own emotional landscape. The aim of DEEP VR is to repeat the virtual experiences enough times to retain "muscle memory" of deep diaphragmatic breathing and develop confidence that they can also "ride through" future stressful experiences. While this technology is relatively new and there not enough studies to show evidence for Bravemind or DEEP VR's long term effectiveness, immersive VR should be considered for patients who don't benefit from traditional exposure (those who have severe impairments or co-occurring problems) and could look to VR as ways to help patients suffering.

Who has access to the 'most engaging' forms of VRET?

According to Kazdin and Blase in "Interventions and Models of Their Delivery to Reduce the Burden of Mental Illness," if were going to solve mental health problems, there is no

[307] Granic, Isabela. "DEEP-VR: A Stunning Virtual Reality Game for Anxiety and Depression (Pt. 5)." It's Your Turn, It's Your Turn, 16 July 2017.

way we can rely on psychotherapy and medications alone[308]. There are over 70% of people with mental illness who receive no treatment from a trained professional and the rates of non-treatment are far higher in low-income countries[309]. Also, standard psychological treatment by a trained clinician is expensive, requires patients to seek public help, demands vulnerability, and requires waiting for available appointments.

VRET apps are often free (or only a few dollars) and significantly less expensive than appointments with a trained therapist. Adjunct head mounts for smartphones can be as low as a $15 Google Cardboard. This device allows users to create an immersive VR experience cheaply, and simply by placing their smartphone at a slight distance away from the lenses, creating a 3D effect by easily moving one's head around to see different angles of the images. Yet, few studies have concluded the efficacy of these apps, and while they may temporarily calm users down, it is unclear how long term the recovery is from using an app. However, significant gains have been made from the past, where VR systems were up to $30,000, much larger and were in limited supply[310]. Therefore, if apps are proven to be efficacious and are included under health care, VR poses the potential to increase access to mental healthcare through VRET apps and fully immersive systems, possibly helping to alleviate the global burden of mental illness.

Threats to Jobs

VR offers a range of technology, from apps to headsets, to fully immersive rooms, and each type of technology requires a different amount of additional human guidance, posing a preliminary threat to jobs for psychologists. VRET apps are user friendly and do not require a psychologist, but these apps have not been proven to be a complete substitute for actual therapy, or proven to be an equivalent to take the place of someone actually engaging in individualized insights and interpretations yet. However, VRET apps could be promising for milder forms of anxiety such as a fear of flying, snakes etc, that wouldn't require a patient to typically see a psychologist anyway and would not threaten any jobs for psychologists. Therefore, there is little evidence to assume VRET apps could a replace a therapist by automating emotional therapy, but these apps could support a large demographic of people are interested in general wellness, and help combat everyday anxiety, stress, burnout, coping, mood management, etc in an informal way.

Furthermore, fully immersive VRET that targets more severe forms of anxiety requires patients to work with a trained clinician, and does not host any evidence that technology could replace the jobs of psychologists but could alter their role. If VRET continues to show efficacy, beyond traditional psychological services, it could impact the curriculum for psychological education and training. Mental health clinicians spend up to 6 years studying to provide efficacious emotional support to patients, but their training could be altered (or reduced in years?) due to the advancements made by VR and potentially be trained in more computer software skills than behavioral therapies. However, therapists are still required to help guide

---

[308] Kazdin, A. E., & Blase, S. L. (2011). Interventions and Models of Their Delivery to Reduce the Burden of Mental Illness: Reply to Commentaries. Perspectives on Psychological Science, 6(5), 507–510. https://doi.org/10.1177/1745691611418241

[309] Thornicroft, G. (2008). Stigma and discrimination limit access to mental health care. Epidemiologia E Psichiatria Sociale, 17(1), 14-19.

[310] Silverman, Dwight. "NASA's Virtual Reality Journey Uses Same Software, Hardware as Gamers." *HoustonChronicle.com*, Houston Chronicle, 12 Mar. 2018.

patients through exposure sequences (Bravemind, DEEP VR), and are asked narrate their fear in present tense back to the clinician, suggesting that prior psychological education would still be important.

Overcoming Stigma of Mental Illness

Therefore, VRET is not a substitute for a trained clinician, but for those who would otherwise not have had access to a mental health clinician or avoided professional help because of stigma, the technology could pose as a helpful resource. There is often a lack of treatment adherence for patients with PTSD and especially for active duty military populations, possibly because of "stigma associated with mental health problems, higher rates of alcohol related problems, and competing responsibilities for preparing for the next deployment."[311] For example, out of 49,425 veterans diagnosed with PTSD, only 9.5% attended nine or more mental health sessions in 15 weeks or less in the first year of diagnosis, but VRET could offer additional anonymity and privacy for patients who would prefer to use this device from the privacy of their own home. Patients do not have to wait for an appointment to expose themselves to a phobia using VR and EBT can be downloaded instantaneously from the convenience from an iPhone. If patients are reluctant to show up to appointments because of stigma, and not for another reason, VR could help increase compliance to treatment.

Problems in Equality

But who gets access to the 'most present' or 'immersive systems? We must consider how this range of immersive VRET systems could influence the equality of access for those who need mental health treatment. VRET apps may increase treatment access to the world's poor, or for those who carry stigma, but there could be consequences from mass producing an emotional and psychological service that may not be as effective as traditional therapy. It seems possible that impoverished communities would have less access to the most efficacious VRET, but if they were supplied with VRET apps over the alternative (no treatment), it does not seem to be doing any harm.

Therefore, VR does not seem to exacerbate a gap in mental health treatment, based on income, race, location etc, but may just not reduce the pre-existing gap, which is hardly an argument not to use it. Fully immersive VR technology that is viewed as more efficacious may only benefit already financially privileged groups, but less immersive VRET apps still leaves room for increased access for people that could benefit from a milder form of anxiety disorder treatment. As a result, VRET does not seem to make the poor any worse off, but could help those who can afford it, making it possible that it may benefit some without hurting others and be considered as an ethical additional form of treatment.

2. VR is Without Real-World Risks (in Response to In Vivo)?

Often, in vivo is not feasible for patients with PTSD like Stacy (war veterans, sexual assault survivors, 9/11 victims) but VR could recreate an authentic and engaging experience, similar to in vivo, but without the same real-word risks. VR uses using a simulated, computer

[311] Seal KH, Maguem S, Cohen B, Gima KS, Metzler TJ, et al. (2010) VA Mental Health Services Utilization in Iraq and Afghanistan Veterans in the First Year of Receiving New Mental Health Diagnoses. Journal of Traumatic Stress 23(1): 5–16

generated simulation that could be a promising solution for patients like Stacy, who needed to change their *perception* of events from 9/11, rather needing to revisit the situation itself. In vivo exposure would be psychologically and physically harmful for patients to relive traumatic experiences, but VR could offer a strong alternative if additional research confirms long term benefits and proves that the additional sensory inputs do not leave users with their own set of side effects (motion sickness, dizziness, nausea, eye strain, etc). Also, we must understand why patients who are suffering from anxiety that *can* be recreated (phobias, fears of heights/public speaking etc) remain averse to in vivo, before we suggest that VR would be promising instead.

Feasibility and Safety

Therefore, the most important hurdle in administering VRET to patients is their consent, and we must consider the feasibility and safety of this technology to persuade them to trying this alternative method of treatment. We must ask patients if they are willing to undergo this form of exposure, and if they find the treatment tolerable. Out of all the literature on VR and anxiety from 'Efficacy of Virtual Reality Exposure Therapy in the Treatment of PTSD: A Systematic Review,' "there was no risk of hidden possible negative results" and remained consistent in the follow up assessments, but it is important to make sure VRET is a treatment that is acceptable for patients and that the additional sensory inputs do not cause significant or permanent distress.[312] There is a strong reason not to think so, but exacerbated symptoms from exposure (for in vivo or simulated) remains a risk factor.

3. VR Increases Imaginal Capacities (in response to Imaginal)?

Thirdly, imaginal exposure may be more controlled than in vivo and pose less real-world risks, but this form of therapy remains limited by the patients' imagination. If the key in recovery is not so much about revisiting reality itself, but for the ability to change their *perception* of reality through their imagination, and a patient's perception of reality is often hindered by their trauma, VRET could increase their imagationary capacities through images. Some patients, like Stacy, express a willingness to imagine their fear using imaginal exposure but are emotionally numb, and are unable to fully engage their emotions and sensory processing. They have difficulty imagining the event or emotionally engaging with the memory and end up telling a flat, emotionless narrative in response.

However, VRET creates a simulated reality that could encourage emotional engagement for patients in EBT that is inadequate from in vivo or imaginal exposure by using additional sensory inputs. VR augments users with visual, auditory and even haptic computer generated experiences and builds on patients imaginal capacites. VR does not require the patient to visualize the feared stimuli on their own, and eliminates a potential barrier for people who struggle with imaging or visualization, and/or are reluctant to think about distressing thoughts. If patients are unable to recover using imaginal exposure because of mental blocks from their imagination, like Stacy, VRET could possibly aid in treatment recovery.

Stacy was unable to emotionally engage with her trauma using imaginal exposure, passively retelling the series of events on 9/11 and continued to suffer from PTSD, but VR

[312] Gonçalves, Raquel, Ana Lúcia Pedrozo, Evandro Silva Freire Coutinho, Ivan Figueira, and Paula Ventura. "Efficacy of virtual reality exposure therapy in the treatment of PTSD: a systematic review." *PloS one* 7, no. 12 (2012): e48469.

helped her engaged more deeply through the visual immersive environment and additional sensory inputs. She was able to recall previously repressed images and details that she was unable to access using her imagination. Computer generated stimuli can capture and guide the patients' attention with less effort than self visualization, and possibly encourage emotional engagement by creating a therapeutic environment for patients like Stacy (who are reluctant to engage in a feared memory). This removes additional variables that could add to the users anxiety and offers potential for a more engaged, and focused treatment session. Therefore, imaginal and in vivo exposure pose limitations for EBT treatment recovery because of their reliance on imagination, or real life, and open an opportunity for VR to be an efficacious mediation between the two standard forms of treatment.

Deception as Risk

"If you assume any rate of improvement at all, games will eventually be indistinguishable from reality," Elon Musk said in a popular podcast with Joe Rogan. "We're most likely in a simulation." In 2003, Nick Bostrom, a University of Oxford philosopher, laid down an 'assault on reality' and argued in a paper that "if there are long-lived technological civilizations in the universe, and if they run computer simulations, there must be a huge number of simulated realities complete with artificial-intelligence inhabitants who may have no idea they're living inside a game — inhabitants like us, perhaps." If computer-loving aliens do exist, but have no physical form, "we are almost certainly living in a computer simulation."[313] Astrophysicist Neil deGrasse Tyson agreed with Bostrom's simulation hypothesis too, and gave a "better than 50-50 odds" that the hypothesis is correct. "I wish I could summon a strong argument against it, but I can find none," he told NBC News.[314]

The simulation hypothesis raises the risk of altering a normal person's perception of reality, but could VRET systems raise *further* risks on a user's perception of reality, for those with mental illness? VR poses a risk that may take advantage of the patient's Anxiety Disorder, and create an environment that deceives them into thinking the simulation is truly real life, or leave them unable to distinguish between reality and fantasy. We should be able to ensure that they will be able to safely distinguish between the two environments, and properly engage in life, and fear stimuli without technology, before engaging in treatment.

In addition, it is important to consider if if users could develop an excessive dependence on VR as a coping mechanism and have further psychological complications like depression etc from the use of VR. For example, if VR therapy is used for loss of a loved one, and if the system visualized the loved one in a realistic way, patients could become dependent on interacting with their image/person and deceived from reality. Likely, doctors/therapists would need to ensure that their therapy regimes don't lead to this unethical outcome but if patients do consent, deception remains temporary (only while using VR), and the technology does not create dependency, VRET does not seem to raise further risks on a user's perception of reality, and lacks a strong enough ethical argument to avoid using it.

---

[313] Bostrom, Nick. "Are we living in a computer simulation?." *The Philosophical Quarterly* 53, no. 211 (2003): 243-255.

[314] Powell, Corey S. "What Is the Simulation Hypothesis? Why Some Think Life Is a Simulated Reality." *NBCNews.com*, NBCUniversal News Group.

4. VR Increases Control and Customization (for Patient and Therapist)?

VR allows patients and therapists greater control over anxiety provoking exposures, and they can use this control to achieve better outcomes by creating a more personalized and customized experience. Users might have difficulty surrendering control to their fear stimulus during 'in vivo,' or struggle to imagine the traumatic event during imaginal exposure, but VR affords greater patient involvement (and subsequent control). The VR environment can be manipulated and explored at the user's own pace, versus relying on the unpredictable nature of real life exposure. Stacy gained control over her trauma by using handheld devices that directed the sequences from her experience on her own terms. This added element of control suggests that VR could increase the patient's' feelings of self efficacy, and allow them to become an active agent of their own experience: important qualities in recovering from any psychological disorder.

VR offers greater control for the therapist too, and can design a highly individualized treatment plan and adjust exposure specific to the patient and their needs. Therapists are able to control the virtual environment using keyboard controls and monitor the patient to prevent negative side effects by viewing the events on another computer monitor, possibly overcoming the limitations held by in vivo. If a patient has a fear of flying, therapists are bound by real life limitations, such as the waiting time and expense for exposing people to flights taking off and landing, but with VRET, therapists can manipulate flying situations suited to their patients individual needs, and maintain control over the stimuli to create the best exposure treatment (taking off and landing a virtual airplane, as many times as needed).

In Stacy's case, she used a VR headset to view a series of sequences tailored to her experience on 9/11, and worked alongside her therapist who monitored her reactions to ensure she was demonstrating the appropriate response. Her therapist viewed the sequences on a secondary computer, and adjusted the situation accordingly to make sure that the fear stimuli were being released at an appropriate rate and controlled for anything unpredictable. If full recovery from anxiety disorder and other mental illness involves recapturing a sense of agency, then it seems essential to understand the roots of this phenomenon and target it through solutions like VR.

Conclusion

Virtual Reality offers an additional form of exposure based therapy that is beginning to show promise as a standard form of treatment for psychological disorders, but it does not come without ethical considerations. VR appears to be a useful technological solution to treat people with anxiety, yet it is important for clinicians to determine why patients are not recovering from standard treatment before considering the benefits to using VR. If patients show enthusiasm to comply, VR could possibly offer advantages over traditional forms of EBT or medication. The four strongest arguments that support VR as a potential efficacious form of therapy is the ability to use additional sensory cues, (potentially aiding in emotional engagement and increasing imaginal capacities), its' lack of unpredictable 'real life' variables (vs in vivo), and its' additional control and customization for the patient and therapist (generating greater feelings of autonomy over their disorder).

However, these arguments for VRET need to be confronted and many ethical considerations should be ruled out before declaring VR as a triumphant tool in mental health

treatment. We must consider the feasibility of this technology, if it is safe and patients are willing to try it. It is important that users are not deceived by the simulation, or that it raises further risks on a user's perception of reality. There is little evidence to think so, but we must ensure that the technology does not trigger an excessive dependence towards VR. In addition, this technology shows promise to increase access to mental health treatment through the proliferation of inexpensive VRET app, but is not a substitute for full treatment recovery.

Fully immersive VR systems may pose less opportunities for increased access to mental health treatment due to its' expensive nature, but may be more efficacious (and engaging) than VRET apps, and this inequality does not seem to make the poor any worse off, just not any better off. These apps could support a larger demographic of people who want to alleviate everyday stress, anxiety etc, and would only help more people who are suffering that wouldn't normally seek professional treatment. Therefore, VR does not seem to threaten the jobs of psychologists and fully immersive VR systems still require the use of a trained clinician. Yet, the nature of psychologists role may change if they begin to use technology in tandem with existing forms of psychotherapy. This could include using more technical skills in their education and adding computer based courses. It's too premature to believe that VRET apps or systems could replace a psychologist, but is important to consider how VRET could be used in tandem with existing forms of treatment and how this technology should be regulated to provide a viable form of treatment for anxiety disorders in the following decades.

Work Cited

Bandelow, B., & Michaelis, S. (2015). Epidemiology of anxiety disorders in the 21st century. Dialogues in clinical neuroscience, 17(3), 327-35.

Bostrom, Nick. "Are we living in a computer simulation?." *The Philosophical Quarterly* 53, no. 211 (2003): 243-255.

Burmester, Alex. “How Do Our Brains Reconstruct the Visual World?” The Conversation, 19 Dec. 2018.

Difede, Joann, and Hunter G. Hoffman. "Virtual Reality Exposure Therapy for World Trade Center Post-Traumatic Stress Disorder: A Case Report." *Cyberpsychology & Behavior* 5, no. 6 (2002): 529-535.

Difede J, Cukor J, Jayasinghe N, Patt I, Jedel S, et al. (2007) Virtual reality exposure therapy for the treatment of posttraumatic stress disorder following September 11, 2001. J Clin Psychiatry68(11): 1639–1647.

Gonçalves, R., Pedrozo, A. L., Coutinho, E. S., Figueira, I., & Ventura, P. (2012). Efficacy of virtual reality exposure therapy in the treatment of PTSD: a systematic review. *PloS one*, *7*(12), e48469. doi:10.1371/journal.pone.0048469

Gorini, Alessandra, and Giuseppe Riva. “Virtual Reality in Anxiety Disorders: the Past and the Future.” Expert Review of Neurotherapeutics, vol. 8, no. 2, 2008, pp. 215–233., doi:10.1586/14737175.8.2.215.

Granic, Isabela. “DEEP-VR: A Stunning Virtual Reality Game for Anxiety and Depression (Pt. 5).” It's Your Turn, It's Your Turn, 16 July 2017.

Jaycox, L.H., Foa, E.B., & Morral, A.R. (1998). Influence of emotional engagement and habituation on exposure therapy for PTSD. Journal of Consulting and Clinical Psychology 66:186–192.

Juan, M. Carmen, and Dennis Joele. “A Comparative Study of the Sense of Presence and Anxiety in an Invisible Marker versus a Marker Augmented Reality System for the Treatment of Phobia towards Small Animals.” *International Journal of Human-Computer Studies*, vol. 69, no. 6, 2011, pp. 440–453.,

Kathmann, Norbert. "Obsessive-compulsive disorder across the life span." (2015): 119-126.

Kazdin, A. E., & Blase, S. L. (2011). Interventions and Models of Their Delivery to Reduce the Burden of Mental Illness: Reply to Commentaries. Perspectives on Psychological Science, 6(5), 507–510. https://doi.org/10.1177/1745691611418241

Kim, Kwanguk, et al. “Effects of Virtual Environment Platforms on Emotional Responses.” *Computer Methods and Programs in Biomedicine*, vol. 113, no. 3, 2014, pp. 882–893., doi:10.1016/j.cmpb.2013.12.024.

Krueger M. Artificial Reality II. Addison Wesley, New York, NY, USA (1991).

Luck, M., & Aylett, R. (2000). Applying artificial intelligence to virtual reality: Intelligent virtual environments. *Applied Artificial Intelligence, 14*(1), 3-32. doi:10.1080/088395100117142

Maples-Keller, J. L., Bunnell, B. E., Kim, S. J., & Rothbaum, B. O. (2017). The Use of Virtual Reality Technology in the Treatment of Anxiety and Other Psychiatric Disorders. *Harvard review of psychiatry*, *25*(3), 103–113. doi:10.1097/HRP.0000000000000138

McLay RN, Wood DP, Webb-Murphy JA, Spira JL, Wiederhold MD, et al. (2011) A Randomized, Controlled Trial of Virtual Reality-Graded Exposure Therapy for Post-Traumatic Stress Disorder in Active Duty Service Members with Combat-Related Post-Traumatic Stress Disorder. Cyberpsychology, Behavior, and Social Network 14(4): 223–229.

Öst LG. One-session treatment for specific phobias. Behav Res Ther. 1989;27:1-7.

Moulds ML, Nixon RD. In vivo flooding for anxiety disorders: proposing its utility in the treatment posttraumatic stress disorder. J Anxiety Disord. 2006;20:498-509.

Powell, Corey S. “What Is the Simulation Hypothesis? Why Some Think Life Is a Simulated Reality.” *NBCNews.com*, NBCUniversal News Group.

Rothbaum, Barbara, et al. “Effectiveness of Computer-Generated (Virtual Reality) Graded Exposure in the Treatment of Acrophobia.” *American Journal of Psychiatry*, vol. 152, no. 4, 1995, pp. 626–628., doi:10.1176/ajp.152.4.626

Roy MJ, Francis J, Friedlander J, Banks-Williams L, Lande RG, et al. (2010) Improvement in cerebral function with treatment of posttraumatic stress disorder. Ann NY Acad Sci 1208: 142–149.

“Understand the Facts.” Anxiety and Depression Association of America, ADAA, adaa.org/understanding-anxiety.

Schuurmans, Josien, et al. “A Randomized, Controlled Trial of the Effectiveness of Cognitive-Behavioral Therapy and Sertraline versus a Waitlist Control Group for Anxiety Disorders in Older Adults.” The American Journal of Geriatric Psychiatry : Official Journal of the American Association for Geriatric Psychiatry, U.S. National Library of Medicine, Mar. 2006.

Seal KH, Maguem S, Cohen B, Gima KS, Metzler TJ, et al. (2010) VA Mental Health Services Utilization in Iraq and Afghanistan Veterans in the First Year of Receiving New Mental Health Diagnoses. Journal of Traumatic Stress 23(1): 5–16

Silverman, Dwight. “NASA's Virtual Reality Journey Uses Same Software, Hardware as Gamers.” *HoustonChronicle.com*, Houston Chronicle, 12 Mar. 2018.

Thornicroft, G. (2008). Stigma and discrimination limit access to mental health care. Epidemiologia E Psichiatria Sociale, 17(1), 14-19. doi:10.1017/S1121189X00002621

Wood D, Murphy J, McLay R, Koffman R, Spira J, et al. (2009) Cost effectiveness of virtual reality graded exposure therapy with physiological monitoring for the treatment of combat related posttraumatic stress disorder. Studies in Health Technology & Informatics 144(7): 223–229.