The Data Renaissance: Analyzing the Disciplinary Effects of Big Data, Artificial Intelligence, and Beyond [Revised Edition]

THE DATA RENAISSANCE: ANALYZING THE DISCIPLINARY EFFECTS OF BIG DATA, ARTIFICIAL INTELLIGENCE, AND BEYOND [REVISED EDITION]

J.J. SYLVIA IV

ABOUBACAR CAMARA; ANA'AYA MCGOWAN MOZELL; BRENDAN SMITH; ELISE TAKEHANA; GLENNDALE BARTOLOME; HENRY CHRISTIANSEN; J.J. SYLVIA IV;

KESHAUNI JOHNSON; LEONORA SHELL; AND SOPHIA MOORE

ROTEL (Remixing Open Textbooks with an Equity Lens) Project

Fitchburg, Massachusetts

ISBN 978-1-964276-49-6 (Print)
ISBN 978-1-964276-48-9 (Ebook)

CONTENTS

Part II. Generative AI in the Classroom and Workspace

Part III. Case Study: "It's Perfect, Four Stars!"

Part IV. Media and Data Literacy

Part V. The American Motion Picture Industry and Big Data

Part VI. Data in Sports Marketing

Part VII. Data in Public Relations, Social Media, and Advertising

Part XII. SOPHIA Discussion Guides

LAND ACKNOWLEDGEMENT STATEMENT FOR THE ROTEL PROJECT

As part of ROTEL Project's mission to support the creation, management, and dissemination of culturally-relevant textbooks, we must acknowledge Indigenous Peoples as the traditional stewards of the land, and the enduring relationship that exists between them and their traditional territories. We acknowledge that the boundaries that created Massachusetts were arbitrary and a product of the settlers. We honor the land on which the Higher Education Institutions of the Commonwealth of Massachusetts are sited as the traditional territory of tribal nations. We acknowledge the painful history of genocide and forced removal from their territory, and other atrocities connected with colonization. We honor and respect the many diverse indigenous people connected to this land on which we gather, and our acknowledgement is one action we can take to correct the stories and practices that erase Indigenous People's history and culture.

Identified Tribes and/or Nations of Massachusetts

Historical Nations

- Mahican
- Mashpee
- Massachuset
- Nauset
- Nipmuc
- Pennacook
- Pocomtuc
- Stockbridge
- Wampanoag

Present-Day Nations and Tribes

- Mashpee Wampanoag Tribe
- Wampanoag Tribe of Gay Head Aquinnah
- Herring Pond Wampanoag Tribe
- Assawompsett-Nemasket Band of Wampanoags
- Pocasset Wampanoag of the Pokanoket Nation
- Pacasset Wampanoag Tribe
- Seaconke Wampanoag Tribe
- Chappaquiddick Tribe of the Wampanoag Indian Nation
- Nipmuc Nation (Bands include the Hassanamisco, Natick)
- Nipmuck Tribal Council of Chaubunagungamaug
- Massachusett Tribe at Ponkapoag

At the time of publication, the links above were all active.

Suggested Readings

Massachusetts Center for Native American Awareness

A guide to Indigenous land acknowledgment

'We are all on Native Land: A conversation about Land Acknowledgements' (YouTube video)

Native-Land.ca | Our home on native land (mapping of native lands)

Beyond territorial acknowledgments – âpihtawikosisân

Your Territorial Acknowledgment Is Not Enough

This land acknowledgement was based on the land acknowledgement of the Digital Commonwealth.

HOW TO USE THIS BOOK

This book embarks on a crucial exploration into how data is wielded across different disciplines, a landscape that is increasingly shaping our modern world. It's essential to note, however, this inquiry is fraught with challenges, particularly because many businesses and organizations guard their data practices as proprietary trade secrets. For instance, platforms like TikTok deliberately shroud their algorithms in mystery, as much of their success hinges on the perception that they've mastered the secret sauce of user engagement.

Our student contributors have diligently navigated these barriers, piecing together an overview of how data is impacting various fields. While we aimed to incorporate DEI perspectives in every chapter, we often found ourselves stymied by the opaque nature of industry practices. The veil of trade secrets not only limits what we can definitively say about data practices in these sectors but also complicates efforts to evaluate these practices from a DEI standpoint.

Despite these challenges, this book fills an essential gap in the current literature. It offers an entry point into the complex interplay of data and industry, providing foundational insights that can spur further inquiry and discussion. However, given the constraints and limitations, this book should not serve as a standalone course text.

For a more rounded educational experience, we recommend supplementing this book with additional resources that focus explicitly on DEI issues. To help in this regard, the appendix includes several Creative Commons licensed readings. We also provide a suggested readings list, carefully curated to complement the perspectives explored in this text and to broaden the DEI lens through which these issues can be examined.

This book is intended to be a living and continually updated document each time the class is taught, rather than a final product. As such, I'd like to briefly explain how it is laid out and how others might use it for their own courses.

Introduction: The introduction is intended to give an historical overview of the project as well as the class in which it was written. I believe transparency is important and the layout of the class may be helpful to other instructors who are teaching a course and want to adopt this textbook. Because of this, it may be of less interest to students themselves.

Chapter 1: Why Care about Data & Society? When I teach this class, it's important to me that I connect with students about what the major issues are related to data and society, and also why I personally care about them. This chapter attempts to bring those approaches together in writing. This approach is not meant to simply share my own accolades. As I discuss in this chapter, I've found that discussing data in the abstract can sometimes cause students to tune out, or make it difficult to connect to the subject. By giving an overview of data through the lens of my own career and experiences, I hope students will be able to see my passion and better understand, in a concrete way, why this topic is important.

Chapter 2 Generative AI in the Classroom and Workplace: This chapter includes a lesson with activities and guiding questions that can be used in class to teach about Large Language Models and other generative AI programs, like ChatGPT. It also features important ethical questions and considerations.

Chapter 3: Case Study: "It's Perfect, Four Stars!": This is a case study written from a first-person perspective by a business professional about their experiences related to data and society. While this is currently the only case in the text, I plan to add more of these in the future, sprinkled throughout the text. This particular case study highlights the voice and experience of a woman business owner.

Chapters 4-10: These chapters were written by students on subjects of their own choosing related to their future career interests. Students in future classes will be encouraged to either expand on existing chapters or write new chapters about career paths that aren't currently represented. Instructors may want to assign only relevant chapters to their students, or allow their students to contribute to the text as I do.

Chapter 11: GenAI hype surrounds us on a daily basis, but so does substantial fear and anxiety. Many worry that such tools will continue to erode critical thinking skills, or remove something that is essentially "human" from the creative process. Others believe that because GenAI tools are trained on the writing and artwork of humans, all use of such tools is a form of intellectual and creative theft. Will the technology continue to improve and eventually achieve sentience? This chapter aims to give an overview of some of these major issues while also demonstrating how to use a variety of GenAI-based tools that might increase the productivity and creativity of professionals.

Chapter 12: SOPHIA Discussion Guides: This chapter features a series of discussion guides that were created by groups of students for public discussions, as part of a partnership with the Society of Philosophers in America. Students in this class could choose to use and modify them to hold their own public discussions, or they could be adapted for in-class discussions.

Chapters 13-14: These two chapters are available via a Creative Commons license and can be assigned to help better fill some of the gaps related to diversity, equity, and inclusion that occur in the above chapters focused on data in specific disciplines. They are written by leading scholars in the field.

Chapter 15: This chapter offers additional recommended reading and viewing, organized by topic. Many of them can be assigned for reading under Fair Use laws or are publicly available multimedia content such as TV shows, podcasts, and documentaries.

PROJECT RATIONALE

I had the privilege of studying with notable posthuman philosophers Rosi Braidotti and Kate Hayles, whose teachings have been instrumental in shaping my worldview. I am a cis-gendered, heterosexual white man with the advantage of tenure at a New England university. I'm acutely aware that these aspects of my identity come with certain privileges and biases.

My education and mentorship in the field of posthumanism—a philosophy that delves into the complex interplay between technology, identity, and the human experience—inform much of my professional and personal life. In simpler terms, my work is driven by a fascination with how humans and technology interact, particularly in shaping our identities. A commitment to feminist ethics provides another critical layer to my approach, enriching my teaching, research, and overall professional practice.

The concept of Becoming is core to my intellectual endeavors. In layman's terms, Becoming means embracing continual change and acknowledging the complex network of connections that make up our lives. This philosophy encourages us to move beyond societal labels or predefined categories. I realize that my identity and perspective are in constant flux, shaped by a myriad of interactions and experiences.

This evolving journey—this process of Becoming—shapes my goal of fostering a learning environment that is inclusive, empathetic, and encourages critical engagement.

I am committed to a continual process of learning and unlearning. The lens through which I see the world isn't fixed; it's fluid, continually molded by an ever-changing landscape of ideas and experiences . What you read here is merely a snapshot, a temporary capture of my current understanding, always subject to future transformation.

I decided to create this book because there is very little existing work that explores how data impacts different disciplines, and almost none written with an undergraduate audience in mind. This stands in stark contrast to the vast resources dedicated to the impacts of data across society more broadly.

As I was developing a new course on the topic of Data & Society that would also meet general education requirements for my campus, I wanted students to understand how data is impacting and re-shaping the specific career fields in which they will be working. One of the things that I've discovered about teaching on topics of data and privacy over the years is that it is very important to make the issues personal for students, as this helps drive a connection and interest with the issues being discussed. One of the ways I wanted to do that in this course was to allow students to better understand how data is being used in their future career fields, while also understanding the ethical challenges associated with those uses.

Unfortunately, there were no resources like this already available, not even in a way that could be cobbled together from multiple different sources. This inspired me to apply for a ROTEL grant that would allow me to collaborate with my students to create this resource as part of our learning experience together. By having students select their own lens through which to write about data, they were able to differentiate their learning and make it personal. One student, in an informal reflection, noted that they enjoyed that they "were allowed to look into what we were interested in through the book chapter."

It was important to me to make this resource available as an OER because it can address an important void in the resources available as classes like this become more common. Further, I believe it's valuable that this book continues to be updated and expanded. My hope is that students in my future classes, as well as other classes that adopt this resource, can see it as an ongoing project that will happily accept the addition of new chapters and updates to those already existing as the field of data studies continues to expand. Ideally, as it continues to grow, students from an even wider variety of fields will be able to find their disciplines represented in these pages and instructors can choose the most relevant sections to assign in their own courses, perhaps in combination with other articles, podcasts, and case studies. For that reason, I've included an additional suggested reading/viewing/listening list at the end of the text that can provide additional content options.

INTRODUCTION

The Data & Society Class: Process and Collaboration

Course Objectives and Structure

This book was created as part of a Data & Society course taught by Dr. J.J. Sylvia IV at Fitchburg State University in Spring 2023. The course was developed as part of a new interdisciplinary major in Digital Media Innovation. Although the major is hosted in the Communications Media department, its classes span nine different disciplines across campus. The course also has general education designations for Civic Learning and Ethical Reasoning. Because the course is open to majors from across campus, it was also tailored to allow students to explore how data is impacting careers and fields related to their own majors and future plans. The course description is as follows:

> This class explores the uses of data in Communications Media, including tailoring professional communication advertising campaigns, green-lighting film productions, creating profitable micro-transaction mechanisms in video

> games, and more. How is data leveraged to form arguments about society, make decisions, and generate profits? Through hands-on projects, students will analyze ethical challenges related to data visualization, algorithms, privacy, citizen and employee surveillance, and more.

The Data & Society class aims to provide students with a comprehensive understanding of the role and impact of data in various industries and sectors, while emphasizing the ethical, social, and cultural implications of data-driven technologies. The course is structured to encourage collaboration, active engagement, and critical thinking, combining lectures, readings, discussions, and hands-on activities to facilitate a dynamic learning experience.

Perhaps most importantly for the current project, the first iteration of this course was designed around the implementation of a Remixing Open Textbooks through an Equity Lens (ROTEL) grant, which supports faculty in their creation of new open education textbooks for academic courses. My approach to this grant was to bring students into the writing process as part of the course requirements. To do this, I assigned core readings addressing pressing issues in the field of data that were front-loaded toward the beginning of the semester. I then invited students to select topics related to their major, planned career, and/or interests. Students worked on this topic throughout the semester by selecting readings for their classmates on the topic, leading a class presentation session on the topic, and drafting and revising their chapter for this text multiple times.

The ROTEL grant also facilitated easy access to a wealth of support across campus, including staff members who were able to visit class and offer support to students throughout the process. I'd like to take this opportunity to offer special thanks to Reneé Fratantonio (Head of Instruction and Information Literacy at Fitchburg State Library), Marilyn Billings (Faculty Advisor & Consultant for a Dept. of Education grant with the MA Dept of Higher Education and Framingham State University), Rachel Graddy (Director of Disability Services at Fitchburg State University), and Meagan Martin (Instructional Designer at Fitchburg State University).

The overall goal for this project is to create a first draft of a text that can continue to be revised and extended by the larger academic community. This will be especially important for a field such as this one, where changes to data practices happen quickly. For example, students in future classes may elect to write new chapters, or update and extend existing chapters based on their interests.

One major limitation of this approach which should be noted is that, especially in a small class such as this one, student interests may align closely. Nearly half of the students who contributed chapters to this first collection were involved in the Game Design major in the Communications Media department. For that reason, we all worked together closely to make sure they addressed different ways that data practices are used broadly within the gaming industry. Ultimately, this means the first iteration of this book is a bit more limited in scope. Nonetheless, I ultimately believe that having students work on a project that they care deeply about is pedagogically more valuable than creating a more topically diverse first draft of this volume.

Adult Learning Connection

One additional element of this grant was to extend opportunities for participation in the project to our local community of adult learners. I did this by teaching a similar course for our Adult Learners of the Fitchburg Area (ALFA) program and offering those students multiple avenues of participation. These courses differ significantly from traditional undergraduate courses in that they do not include grades or traditional assignments, though reading lists are a common element. For this reason, ALFA students were invited to participate in the project in a few different, entirely optional ways, which included submitting a chapter of their own, helping with editing, or mentoring undergraduate students and providing feedback on their work. Ultimately, two of these ALFA students, Kevin and Carol Smith, chose to mentor students in my undergraduate class and visited several times to provide feedback on work-in-progress.

Student Contributions and Chapter Development

Here, I'd like to take a moment to fully outline the process that was used for the development of the chapters included in the text by students, especially in case other courses may like to adopt or amend this process.

Week 3: Students were asked to select a general topic early in the semester, by the end of week three. Their choice here was not yet binding, but meant to provide a guiding framework for their next step, as they developed a larger proposal for their writing. This week featured multiple guests visiting the course. Marilyn Billings gave a presentation that covered the overall goals of the ROTEL grant and why we are creating OER textbooks. She also helped students develop an understanding of the creative commons licenses available and we had a discussion about the type of license we wanted to assign to our project. Reneé Fratantonio also gave a demonstration on how to use library resources to complete research on their topic. Students were given time in class to begin researching their chosen topic and ask for help or guidance from myself and Fratantonio. ALFA mentors Kevin and Carol Smith were also in class on this day and had conversations with all the students in the course. In addition to any feedback the ALFA mentors gave, the challenge of putting their idea into words and talking through it with someone not directly involved in the class was itself valuable to students in the process of selecting their topic.

Week 6: Students were next tasked with writing an approximately 150-300 word proposal for their topic, requiring them to complete additional research on their proposed topic to make sure it was viable. ALFA mentors attended class again on this day to hear the revised and extended proposals and offer feedback. In this class session, students rotated through meetings with both the ALFA mentors and me to workshop these proposals and prepare for the next steps of writing a full chapter. Students also signed up for the day in the semester where they would lead a class session on their chosen topic.

Week 9: The first full draft of the chapter was due during the ninth week of the course, with the expectation that it may still be a bit rough around the edges as students were continuing to learn more about their chosen topic. For this draft, I provided big-picture feedback on the chapters-in-progress. This included suggesting parts of the topic that perhaps were not addressed or needed expansion. I also gave feedback on how students could more deeply address diversity and ethics within their topic. Rachel Graddy and Meagan Martin visited class on this day to discuss accessibility considerations for the writing process.

Week 14: The second draft of the was chapter due this week, and this was intended to be a complete and polished draft that students would consider ready for publication.

Week 15: Between weeks fourteen and fifteen, students all peer reviewed one another's work using Google Drive commenting and suggesting tools. I also participated in this review process, leaving extensive feedback that included minor issues such as grammar as well as major suggestions for revisions.

Week 17: The final draft of these chapters was due during week seventeen, which was the scheduled final exam period for the class. No exam was given during this period, but students could attend with final questions about the project at this time.

Finally, I should note that some further editing was completed by me on these final drafts. However, in this round of editing I only focused on minor edits aimed at clarity and did not make any structural or thematic changes to the final product created by students. In short, students were provided with significant on-campus support, multiple rounds of iterative feedback, and opportunities to fine tune through three drafts of the chapters. Ultimately, this assignment was challenging for students, as the majority of them had not previously encountered any similar coursework about the implications of data on society, and were therefore exploring an entirely new subject area. The majority of students were also freshman or sophomores.

The Importance of Understanding the Implications of Data on Society

There were important tradeoffs in approaching the course in the manner described above and in letting students help develop the topics addressed in the class. Most significantly, this meant that I was selecting only about half of the overall content of the class in terms of the topics and readings assigned, while the rest was ultimately assigned through the decisions made by students. Therefore, I tried to highlight the major societal issues related to data. Briefly, I explored the following topics during the class:

A Brief History of Information and Big Data: Understanding the historical context of data development and the emergence of big data helps students appreciate the evolution of data-driven technologies and their impact on society.

Artificial Intelligence: Exploring the development and applications of artificial intelligence (AI) provides insights into the ways AI has revolutionized various fields and the ethical considerations that arise from its use.

Data Bias and Algorithms: Examining issues of data bias and algorithmic fairness is essential for understanding how data-driven technologies can unintentionally perpetuate existing biases and discriminatory practices. Students explore a range of readings on this topic, such as works by Jill Walker

Rettberg, Kate Crawford, Catherine D'Ignazio, Lauren F. Klein, and Safiya Noble.

Data Ethics: Delving into the ethical considerations surrounding data collection, use, and dissemination helps students develop a responsible and conscientious approach to data-driven practices. Resources such as "An Introduction to Data Ethics" by Shannon Vallor and William J. Rewak provide valuable guidance.

Quantified Self and Data Visualization: Investigating the quantified self movement and data visualization techniques enables students to explore how data shapes our understanding of ourselves and the world around us. Students engage with readings from Jill Walker Rettberg, Claudo Minca, Maartje Roelofsen, and the Tableau Public Blog.

Analyzing Social Media: Studying social network analysis methods allows students to examine the ways in which data informs our understanding of online interactions and social networks. Resources such as the works of Gruzd, Paulin, and Haythornthwaite, along with Netlytic Video Tutorials, provide a foundation for this exploration.

The Emergence of AI and ChatGPT in the Course

Finally, I believe it would be remiss not to address the historical significance of the rise of OpenAI's ChatGPT and other AI-based tools during the semester this course occurred.

ChatGPT had officially been released in November of 2022, shortly before the course began, therefore I was, to some degree, able to anticipate this change and include related readings on the syllabus. However, the rate at which the tool was updated and the speed with which it was adopted felt overwhelming at times and required hours of attention on a weekly basis to keep up with the ongoing developments.

Every few weeks in class we would check-in on these ongoing developments, discussing especially the ethical issues connected with the technology. As part of our assigned lesson, students also spent time in class working with ChatGPT to better understand its affordances and limitations. One theme that emerged from our in-class discussions was that, at least in its current iteration, ChatGPT was very helpful as a brainstorming tool and to edit or explain existing text but was less helpful in generating specialized writing and essays, especially if they required the use of citations. Going further, one assignment in the class that students could choose from among a list was to test how ChatGPT performed on an assignment they had in another course. All students who completed that assignment reported that ChatGPT was not able to do a satisfactory job completing the assignment they chose.

Because this technology was emerging so quickly, the policy I put in place for the use of AI this semester was simply one that required transparency. I asked that students note any time they used generative AI tools along with how they were used.

This is a policy I plan to revisit after reflection on how it went this semester.

In an effort to promote transparency and ethical use of AI, students who incorporated insights or assistance from ChatGPT in their chapters were required to acknowledge its use. This practice ensures that readers are aware of the role of AI in the development of the content and fosters open dialogue about the implications of AI in research and writing. I used ChatGPT for brainstorming and editing purposes in my writing for this text, as a way to further experiment with the tool. However, I did not use it to entirely generate any portions of text.

PART I
WHY CARE ABOUT DATA AND SOCIETY?

Chapter Written by J.J. Sylvia, Ph.D.

> *"I'm not worried about privacy because I haven't done anything wrong."*
>
> – Most People

Learning Objectives

- Explain the ethical challenges and societal implications of big data, including issues of privacy, trust, and potential for misuse.

- Understand the various approaches to regulating big data and why traditional methods like notice and consent or anonymization are increasingly insufficient.
- Critically discuss the intersection of race, gender, and capitalism in the realm of data science, recognizing how biases can be built into algorithms and data sets

INTRODUCTION

As a tenured professor deeply immersed in the confluence of digital media and posthuman philosophy, my life's work has largely revolved around deciphering the intricate web of technology, identity, and human experiences. Prior to becoming an academic, I worked in the ecommerce sector for two decades, and spent five years working for a nonprofit organization that helped K-12 schools better integrate educational technology. The impetus for this chapter comes from a profoundly personal place—a purposeful sense of self that draws from both my academic background, my professional background, and inherent interests in the subject at hand.

The journey we'll undertake in the following pages isn't just a scholarly expedition; it's also an exploration of my own evolving understanding of how technology can both empower and marginalize, illuminate and obfuscate. In this sense, the chapter serves as a dual lens: one that presents a specific subject matter through the filter of academic rigor, and another that invites you to understand how my own experiences and intellectual journeys have shaped this presentation.

My hope is that the ensuing discussions will not only add to your knowledge base but will also inspire you to consider your own positionality—your unique vantage point formed by your experiences, background, and education. Just as I have connected my own life story to this area of study, I encourage you to discover your own connections, contradictions, and curiosities as we delve deeper into the complexities of this intriguing subject.

As I've taught about the impacts of big data and artificial intelligence (AI) over the years, I find myself frequently running headfirst into one formulation or another of the above quote, which I've obviously made up and not cited exactly. Or perhaps it's more accurate to say I've been trained on a large set of data that consists of responses to concerns about privacy, processed those through my neural network, and generated some predictive text that looks a lot like what most people say – much like any good large language model (LLM) would do as part of a generative AI process. Either way, developing an approach to teaching about data that cuts through the apathy associated with this quote, or one like it, has become a central focus of my pedagogy. Why exactly should we spend our precious time on this planet thinking about or even *caring* about ideas as abstract and hard to regulate as these?

As it turns out, there are quite a few good reasons. The challenge is these reasons are buried in layers of legal and bureaucratic jargon that, frankly, make it all sound quite boring. Comedian John Oliver described this best when discussing the intricacies of net neutrality and cable companies on his show *Last Week Tonight:*

> Oh my god! How are you still so dull? And that's the problem. The cable companies have figured out the great truth of America. If you want to do something evil, put it inside something boring. Apple could put the entire text of Mein Kampf inside the iTunes user agreement and you'd just go, "Agree, agree, agree. What? Agree, agree." (Oliver, 2014)

Oliver goes on to distill the issue of net neutrality, explaining it in detail while also making it funny. While I would love to be able to do something like that in every class session that I teach, the amount of content I have to produce every semester while teaching four courses is far greater than the amount of content someone like Oliver produces for his show, and he has an entire team of writers helping him. Nonetheless, I've worked hard over the years to find ways to make the big picture questions about data and society both personal and interesting to students. Let's explore why this matters.

A CONSTITUTIONAL RIGHT TO PRIVACY?

In the United States, the right to privacy moved into the spotlight as part of the controversial 2022 Supreme Court decision in Dobbs v. Jackson Women's Health Organization. This decision removed federal protections for abortion rights, instead deferring the right to legislate abortion to individual states. As monumental and disruptive as that particular decision was, the fallout from legal precedent it overturned to do so is arguably even larger. The syllabus that gives an overview of the case explains:

> As to precedent, citing a broad array of cases, the Court found support for a constitutional "right of personal privacy." Id., at 152. But Roe conflated the right to shield information from disclosure and the right to make and implement important personal decisions without governmental interference. (Dobbs v. Jackson Women's Health Organization, Syllabus, 2022, p. 5)

Let's break down what this means. While the Court did not eliminate the constitutional right of personal privacy, it argued that the Roe v. Wade decision, which originally legalized abortion at the federal level, misconstrued how the right to privacy actually works. Roe v. Wade made the argument that the right to privacy means that citizens have the right *both* to shield private information from disclosure to authorities *and* use that right to make personal decisions without government interference. In other words, the right to legally obtain an abortion was based on the constitutional right to privacy. The Dobbs v. Jackson Women's Health Organization breaks that link, arguing that the right to shield disclosure of an action does not confer the right to take that action. Said another way, although one has the right not to disclose information about whether they've had an abortion, that right does not make the act of getting an abortion legal.

Although this may appear to be a minor distinction, it potentially disrupts the entire foundation of the right to privacy in the U.S. The majority opinion said this ruling should not affect other cases on which legal rights were tied to the right to privacy. However, in a concurring opinion, Justice Thomas Clarence argued just the opposite:

> For that reason, in future cases, we should reconsider all of this Court's substantive due process precedents, including Griswold, Lawrence, and Obergefell. Because any substantive due process decision is "demonstrably erroneous," Ramos v. Louisiana, 590 U. S. ___, ___ (2020) (THOMAS, J., concurring in judgment) (slip op., at 7), we have a duty to "correct the error" established in those precedents, Gamble v. United States, 587 U. S. ___, ___ (2019) (THOMAS, J., concurring) (slip op., at 9). (Dobbs

> v. Jackson Women's Health Organization, 2022, Thomas, J., concurring, p. 3)

Here, Thomas is specifically arguing that in light of the Court's decision in Dobbs v. Jackson Women's Health Organization, the court should revisit other cases that used the same precedent as Roe v. Wade and correct the error in those cases. What cases does he mention? The 1965 Griswold v. Connecticut case predated Roe v. Wade and established a constitutional right to privacy, recognizing that married couples have the right to use contraceptives. The 2003 Lawrence v. Texas Supreme Court case declared laws criminalizing consensual same-sex sexual activity unconstitutional, affirming the right to privacy and striking down sodomy laws in the United States. The 2015 Obergefell v. Hodges Supreme Court case legalized same-sex marriage nationwide in the United States, recognizing it as a fundamental right protected by the Constitution. In short, Thomas is recommending that the Court revisit the cases that protected the rights to use contraception, to perform consensual same-sex activity, and same-sex marriage and "correct the error" that was made in those decisions.

Let's revisit that quote at the beginning of the chapter in light of this discussion:

> "I'm not worried about privacy because I haven't done anything wrong."
> – Most People

Rather than saying you aren't worried about privacy because you haven't done anything wrong, you should instead understand that in the United States, at least, what counts as right or wrong under the law has long been guided by the constitutional protections of privacy. But we are now living in a shifting landscape where these protections will no longer stand on firm ground. Let's consider one very personal example of data and privacy that has shifted in light of the Dobbs decision.

Many women have long tracked their menstrual cycles for a wide variety of reasons, including, but not limited to better understanding their health, as a form of birth control, as a way to increase the likelihood of conception, for medical reasons, and to look for signs of menopause. A plethora of cell phone apps are available that can help track this information. None of the activities listed above are illegal, so it may be easy to believe there's no need to be concerned about privacy in this case. However, if one lives in a state where abortions are no longer legal post-Dobbs, this data can potentially be collected and used as evidence of abortion if there are irregularities (which can occur naturally) in menstruation cycles. Efforts to protect period-tracking app data specifically have thus far failed (Moomaw, 2023). It's important to note, though, that digital evidence can be collected and used against those who seek abortions from sources far beyond period-tracking apps, including wearable technology, internet-connected household appliances, purchase history, routine data gathering by government agencies, and social media usage (Conti-Cook, 2020). All of these sources of data travel outside of one's home because they travel over the internet, therefore they are not protected by the remaining right to privacy.

Without the legal protection of privacy, many of our previously guaranteed rights, including whom we marry, whom we engage with in sexual activity, and whether or not we have children are either no longer legal already or may not be in the near future. It doesn't get much more personal than that.

LITTLE BROTHER

George Orwell's dystopian novel coined the term "Big Brother" for overly intrusive governments that use surveillance to erode privacy. However, in the era of big data, the use of our data in other areas of society should also raise concerns, as we now live in a culture of algorithms. I have elsewhere called the private companies that use our data, as opposed to the government, Little Brother (Sylvia IV, 2016a). Here, too, I have learned that if one is going to care about how data is being used, the consequences of it need to be felt personally. Let me briefly walk through how this approach has developed as part of my professional work on topics related to big data.

An interactive H5P element has been excluded from this version of the text. You can view it online here:
https://rotel.pressbooks.pub/datarenaissance/?p=38#h5p-1

Podcast: Living in a Culture of Algorithms

Episode Summary:

danah boyd weaves together her work on youth, privacy, and data-driven technologies, to examine the complicated social and cultural dynamics underpinning social media, the messiness of "big data," and the problematic implications of using algorithms designed for one problem to address societal issues without accounting for unintended consequences.

Aperveillance

My goal is to make questions about data come alive, using creative and/or artistic practices that allow us to understand the ethical challenges presented by big data in new ways. My first attempt at this was a project titled Aperveillance.

Figure 1. Aperveillance project displayed on the media wall at Hunt Library, North Carolina State University. The image shows a grid of multiple live webcam images overlaid by text noting the crimes committed in Raleigh, NC the previous day.

On my website, I describe the project in this way:

While teaching a special topics course (COM/ENG 395) on Big Data and the Rhetoric of Information, I asked my students to create data visualizations with either Tableau or P5.JS, and I joined them in creating my own project for the assignment. Although much of the concern in my field surrounds issues related to surveillance, I thought it would be interesting to think about the types of watching that are now becoming increasingly possible with open data. Thus, I coined the term aperveillance for my project, which derives from the Latin "aper" meaning open, and "veiler" meaning to watch.

This project uses webcam images that are publicly available in North Carolina, primarily around Raleigh, but including other areas of the state… It also uses Raleigh's open crime data to randomly include information about the previous day's crimes juxtaposed on top of the webcam images. This is intended to provoke questions about the type of watching we as citizens are able to do with open data on the web. (Sylvia IV, 2016b).

This project was displayed as part of a Code+Art exhibit at the North Carolina State University library and at a local conference. However, on a whim, I made one last-minute tweak to the project, not visible above, which ended up being the most interesting part of the project. For anyone who was viewing the project on a device that had a camera attached or built-in to it, I grabbed an image from that local camera and mixed it randomly into the grid of local webcam photos. This was by far the aspect of the project that generated the most interest while the project was being displayed. To my surprise, audience members posed questions that reflected significant concerns about their privacy once they saw themselves displayed in the data. This served as a moment of inspiration that would lead to my next project.

I should also note that, although this project was viewable live on the web, the images taken from the viewer's webcam were only ever displayed on the local device on which one was viewing the project. They were never displayed to anyone else via the internet or saved or archived in any way. But, the concerned reactions by viewers helped me better understand that even creative projects like this one, which used local data, would only have the impact I was seeking if the impact was felt in a truly personal way.

BECOMING DATA

This insight led directly to my next interactive project. While teaching at Fitchburg State University, I secured a small grant that allowed me to hire two students to help write and code this project. Much like the Aperveillance project, this project also relies on a bit of trickery, in which the program itself acts as if it's completing a data analysis which does not actually occur. However, it's important to understand that in both of these cases, the trick could actually be implemented for real, but is not done so in order to protect privacy. In other words, the Aperveillance project could have displayed the images from the local webcam live to everyone via the internet, and even saved them. And the Becoming Data project could have actually completed the data analysis that it fakes. However, I'm not personally interested in violating anyone's privacy – I simply want them to experience what such a violation feels like at a personal level that is not possible when discussing data abstractly.

Becoming Data uses a Microsoft Kinect and the Processing programming language to create an augmented reality interface in which users interact with the screen in order to start an analysis of their own data. It includes a fairly annoying terms of service agreement that must be navigated and accepted. Then the system acts as if it is performing the following tasks, with a percentage completion bar and animations for each:

- Facial Recognition Analysis
- Pinging cell phone
- Sentiment Analysis
- Social Network Analysis
- Accessing Credit Data

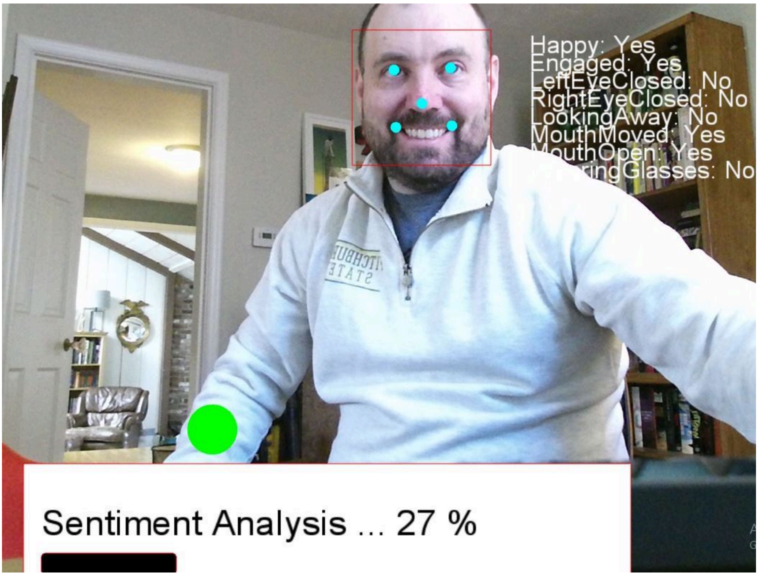


Figure 2. Becoming Data screenshot. The image shows the Microsoft Kinect in the middle of performing a sentiment analysis. The face in the image is surrounded by a red box with cyan dots marking the eyes, nose, and corners of the mouth. Text on the image indicates the man is happy, engaged, has both eyes open, is not looking away from the camera, moved his mouth, has his mouth open and is not wearing glasses.

Next, the program shares the following results screen-by-screen, with a simple random number inserted where each bold number is included below:

1. Total Time You Spent Reading License Agreement: [Actual number recorded] seconds. This agreement was approximately 2,500 words. The average time to read this

amount of text is between 8 and 20 minutes. What did you miss?

2. Analysis of recent food purchases indicates that you may be depressed! Social media channels will now feature ads for meditation apps and online therapy services **28%** more frequently. If you click one of these, you may start receiving ads suggesting you are bi-polar.

3. Location-based data collected from your cell phone indicates that you have visited the gym less frequently than the national average. Your health insurance rates will rise **7%** this year.

4. Based on an analysis of your cell phone battery, ride sharing services will increase your fare by about **33**%.

5. Car dealerships can access your recent search history. Based on an analysis of your recent searches, if you were to visit a dealer today, they might offer you a loan with an interest rate that is **2%** higher.

6. Using a 2015 patent, the creditworthiness of your friends across social media sites has been analyzed. Unfortunately, some of your friends have low scores. As some new credit card companies take this into account, your credit score could drop by as many as **171** points.

7. An analysis of all of your social media posts reveal that

your posts contain **76** bad words and **115** mentions of alcohol or drugs. These may or may not be problematic in context, but they have been archived and will be reported on your next employment background check.

8. The web browser you use most frequently on your phone has been correlated with an increased likelihood to leave a job sooner than other employees. Approximately **21%** of employers will not even consider your application due to the browser you use.

9. Your recent social media feeds show **65**% more ads with negative sentiments recently. This increase may be due to experiments run by the company or an influx of memes by Russian-backed ads. However, this change in your feed means you are **33**% less likely to vote in the next election.

10. A stress analysis of your face indicates that you are **57**% unlikely to be placated by a customer service representative. If you call for customer service now you will be routed to an operator who has been specially trained to handle difficult customers. They will be unwilling to meet your requests.

This project was completed days before the COVID-19 pandemic caused schools across the country to close in early 2020, so it has had limited opportunities for sharing. However, it has been featured on my campus as part of the Speaker's Series for the Center for Teaching and Learning and as part of classes that I taught once students returned to in-person learning.

Participants and the audience members viewing the interaction have had very strong reactions, which usually first focus on how unfair a particular result is, followed by questioning if the results are real. As I noted above, the actual results presented are fake, but all of the situations shared in the results are based on real-world patents or applications of data use. This is the best way I have found to date to make the effects of Little Brother – the use of your data by corporations – feel real and feel concerning. Our personal data can be collected and used in ways that are detrimental to not only our wallets but also our very livelihood.

Does this make you care?

A/B AND MULTIVARIATE TESTING

I have further argued that even without access to *personal* data, massive amounts of data cause an ethical problem related to manipulation. This manipulation is related to the ethical implications of A/B and multivariate e-commerce optimization testing (Sylvia IV, 2010). These techniques, which allow e-commerce websites to test different versions of a page to improve outcomes like sales or reduce abandoned shopping carts, might seem innocuous. However, I believe there's more to consider.

I've been involved in owning or managing e-commerce websites for 20 years, and I first became aware of the issues surrounding this sort of testing in the first decade of my career. It was around this time that Google's free Web Analytics software launched and was available for free to the general public. This allowed virtually any site that wanted it to run these A/B and multivariate tests and collect data on them. I saw the power of these tools first hand as I integrated them into the site I was managing. It was witnessing this power that first raised my ethical concerns. This is part of what brought me back to school to pursue my master's degree, and later my doctorate.

I examined these practices through various ethical lenses and I've found that they can lead to manipulative site design. The goal is to subtly encourage consumers to spend more. Although another viewpoint might see these practices as aiding consumers—making websites more user-friendly or easier to navigate—I think the reality is more complex. The goal, much like the field of advertising in general, is to create new desires to purchase products you don't actually need. But this iterative process lets websites get really good, really quickly at persuading you in ways that are not at all transparent. How could you possibly imagine that the color of the checkout button on your favorite website makes you more likely to actually complete a purchase unless you've studied web design or communication theory?

The primary issue that I took at the time with these practices was that the *why* didn't matter. Why does a certain size and color button make people spend more? Why does a certain shade of blue make people more likely to click a link? This type of testing cannot answer that question. As we transitioned into the age of big data, that problem has only become more pronounced. Big data is very good at making correlations between things, but not able to explain why those correlations exist.

And this brings us face-to-face with the difficult theoretical questions we must all face in the age of big data.

BIG DATA

The Five V’s

What we refer to as big data is typically defined through the five v’s definition: volume, velocity, variety, value, veracity. Put as simply as possible, the five v’s include a massive amount of different types of data that are being collected with increasing frequency from multiple sources. Outputs are providing great value to the organizations that can make use of it, while presenting significant challenges if one needs to determine the accuracy or truth of the content represented by such data.

Where does all of this data come from?

Early on, most of it was generated by human actions, through the data we leave as we browse the internet and use devices with sensors built into them, from our cell phones and smart watches to the thermostats and doorbells in our houses. But the low cost and huge amounts of data generated by sensors has led to their implementation into smart cities, shipping processes, and beyond in ways that allow them to collect data on the world that goes beyond the human. For example, most international shipping now uses RFID tags to collect information and monitor shipments. Just how cheap are all of these censors? According to DuBravac (2015), a typical smartphone in 2015 could have all of the following sensors for an additional $5.00 in manufacturing costs: proximity, ambient light, accelerometer, gyroscope, magnetometer, ambient sound, barometer, temperature/humidity, and M7 motion. Check out the documentary below for an overview of how big data is being used:

Documentary: The Human Face of Big Data

Documentary: The Human Face of Big Data on Vimeo

But this leads to yet another question: why do we so willingly give up all of this data for free to corporations that use it to manipulate us and increase their profits?

Access to Data: Weapons of Math Destruction

Although he has since fallen into significant controversy because of his political views, journalist Glenn Greenwald (2014) spoke clearly about this challenge in his TEDGlobal talk. Greenwald was one of the journalists who helped NSA whistleblower Edward Snowden publish his story about the way that the U.S. government was abusing the U.S. Patriot Act to illegally collect information on U.S. citizens. In that speech, Greenwald notes that we do seem to intuitively care about privacy. For example, if someone were to ask us for our email address and password, we very likely wouldn't share that information, even with close friends.

And yet, we give up the contents of our personal email to corporations like Google and the details of our social lives and personal messages to companies like Meta, which owns Messenger, Instagram, and What's App. One possible reason we might feel comfortable sharing this information is because we trust these companies. For many, this was explicitly true when it came to Google, at least for many years. However, not everyone trusts technology firms in the same way:

> When we consider the race of our respondents, white individuals (the baseline/omitted category in our model) are the racial group that is least confident in the three tech companies, save for respondents who identified as multi-racial or as some race other than our main four groupings. Interestingly, there doesn't seem to be a meaningful difference between Asian, Hispanic, or Black respondents. (Kates et al., 2023, para. 17)

In short, Asian, Hispanic, and Black people trust technology firms such as Google more than white individuals. This means that they are more likely to share personal data and less likely to consider the negative impacts that can stem from that sharing. Further, any education past high school led to a decrease in trust. Gender showed some difference in trust levels, but was relatively small or had a small enough sample size so as to decrease the overall statistical significance of the results:

> ...respondents identifying as female [were] slightly more confident than males in our tech companies, but the substantive magnitude of this difference is quite small. Those identifying as either non-binary or neither male nor female, however, are vastly less confident, though our results only reach significance at the 0.10 level, given the paucity of such respondents in our panel. (Kates et al., 2023, para. 19)

Until they eliminated it in 2018, Google's company motto was "Do No Evil." If you've been paying attention to the world of technology, you can already see where this story is heading. Google has been the subject of antitrust investigations, security vulnerabilities that left personal data accessible, and fears of search-induced filter bubbles that may have helped sway political elections. Many of those who trusted Google with their intimate and personal data in the early 2000s no longer do so. Although people have lost trust in all institutions, their trust in technology companies, in particular, decreased the most drastically between 2018 and 2021. Notably, this was true across every sociodemographic category analyzed (Kates et al., 2023). Overall, trust in technology companies has decreased for everyone.

Cathy O'Neil describes the use of this data in the form of algorithms, "weapons of math destruction." In the podcast below, she explains how this works and how it magnifies inequality in our society.

Podcast: Weapons of Math Destruction with Cathy O'Neil

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://rotel.pressbooks.pub/datarenaissance/?p=45#h5p-2

Data & Society: Weapons of Math Destruction

Episode Summary:

Tracing her experiences as a mathematician and data scientist working in academia, finance, and advertising, Cathy O'Neil will walk us through what she has learned about the pervasive, opaque, and unaccountable mathematical models that regulate our lives, micromanage our economy, and shape our behavior. Cathy will examine how statistical models often pose as neutral mathematical tools, lending a veneer of objectivity to decisions that can severely harm people at critical life moments.

Cathy will also share her concerns around how these models are trained, optimized, and operated at scale in ways that she deems to be arbitrary and statistically unsound and can lead to pernicious feedback loops that reinforce and magnify inequality in our society, rather than rooting it out. She will also suggest solutions and possibilities for building mathematical models that could lead to greater fairness and less harm and suffering.

However, even if that's not your personal experience, or even if there is a corporation you trust implicitly, no corporation lasts forever. And when that company is sold or dissolved, its assets are often transferred elsewhere, possibly to much less trustworthy owners. Although we may be aware of that possibility in the abstract, I would like to share a case study about how the implications of this process impacted me.

Livejournal Case Study

This reality became personal for me in 2019, as I was researching Russia's internet policies as part of an article I was writing with a colleague about Russia's interference via social media in the 2016 U.S. presidential election. While doing that research, I discovered that the social media site LiveJournal, which had been popular in the very early 2000s, had not only been sold to Russian oligarchs, but all of their servers were physically moved to Russia. Why did this matter so much to me?

A short history of LiveJournal can make this clearer. Its origin story is somewhat similar to that of Facebook in that it was launched out of the college dorm room of its creator Brad Fitzpatrick in 1999. I had already been blogging for several years by the time the site began to gain popularity. In fact, as best as I can tell, I very likely had one of the first one hundred blogs ever published on the internet when I launched mine as a high school sophomore in 1998. My friends and I competed with one another to release new and more creative features for our blogs. But this interest in the software behind the blog gave way to a more sustained interest in the content of the blogs. Fitzpartick's new site also allowed the creation of friends lists, which meant that rather than taking the time to visit each of our blogs separately, we could all sign up for accounts and have the most recent updates appear in one feed, in chronological order. This is standard today, but was a huge leap forward when it was created.

This means I was using LiveJournal as I transitioned from high school to college. This can be a highly emotionally turbulent time, as you may be aware. Many of us who used LiveJournal at the time would write very long and very personal entries. Of course, it also had quite advanced security features, meaning you could create customizable lists that determined who could see each specific post. While this particular feature still exists on some platforms today, it has yet to be replicated in such an intricate way as LiveJournal allows. I also wasn't alone in my usage of LiveJournal. It peaked at over 2.6 million active users within a 90-day period in 2005.

These filters, and an implicit trust in Fitzpatrick, gave me confidence to write about very personal things online. Because Fitzpatrick also posted in his own journal, it felt very much as if I knew him personally, though my later study of communication theory would reveal that this was really only a parasocial relationship. As my life continued to evolve, I slowly stopped using the platform, and hadn't thought about it in some years until the day I stumbled across the news of its move to Russia. Why does this all matter to me?

The short version is that LiveJournal was sold a few times over the years before it ultimately ended up in Russia. The key here is that Russia's laws allow the government to access any information on servers located in their country, without the kind of strong protections like the need for a warrant that are in place in the United States. Does it really matter that the Russian government now has easy access to all of my old private, password protected writing? Probably not. I haven't revisited the volumes of writing I did there in well over a decade, but as far as I remember, there was nothing truly egregious that I ever posted. But at minimum, the detailed musings of myself as a teenager could certainly be embarrassing and almost definitely cringeworthy to the version of me that is now a tenured professor. The types of things people posted about then weren't as curated and glossy as they are today. We would post about things we clearly coach people not to post on the internet today.

As my professional research has progressed into criticisms of Russia and their impacts on democracies around the globe, a small voice in my head can't help but wonder if there's something somewhere in all of that writing that could be used against me, especially if it were taken out of context. Russia is well known to operate blackmail schemes.

And to think, all of that worry because the teenage version of me placed so much trust in Brad Fitzpatrick. And yet, we know that others are at much greater risk. In the 2016 election, Russian troll factories specifically targeted Black and Latinx U.S. voters on social media, actively dissuading them from voting at all as a way to bolster Donald Trump's success in the election. Since then, their methods have gotten even more complex. For example, they have set up fake sites designed to look like they offer help for those struggling with their sexual identity and how or whether to share it with friends and family. The Russian trolls then use those conversations to blackmail the participants into taking actions that advance Russian goals (Sylvia and Moody, 2019).

Racial Capitalism

As we saw in the last section, everyone is at risk when our personal stories and data become entangled with websites, even those we may initially trust. However, that risk is not evenly dispersed, as marginalized people are almost always the most significantly impacted by the challenges our society faces related to data and algorithms. These challenges have many layers, but they begin at the very beginning of our technology, during the coding process itself. If we're discussing Little Brother, corporations who use data, then connections between capitalism and racism are a necessary piece of the puzzle needed to untangle this story.

Sometimes, these implicit biases emerge because the technology is created predominantly by white people who only test the code on other white people or use data sets that don't reflect diverse people and/or skin tones. Why does this happen? The technology workforce is overwhelmingly white. For example, only 4% of Google's workforce is Black, Black people represent only 1% of tech projects that receive venture funding (Russonello, 2019). The following documentary, *Coded Bias*, explores these challenges:

Documentary: Coded Bias by PBS

PBS: Coded Bias

In an increasingly data-driven, automated world, the question of how to protect individuals' civil liberties in the face of artificial intelligence looms larger by the day. Coded Bias follows M.I.T. Media Lab computer scientist Joy Buolamwini, along with data scientists, mathematicians, and watchdog groups from all over the world, as they fight to expose the discrimination within algorithms now prevalent across all spheres of daily life.

While conducting research on facial recognition technologies at the M.I.T. Media Lab, Buolamwini, a "poet of code," made the startling discovery that some algorithms could not detect dark-skinned faces or classify women with accuracy. This led to the harrowing realization that the very machine-learning algorithms intended to avoid prejudice are only as unbiased as the humans and historical data programming them.

Coded Bias documents the dramatic journey that follows, from discovery to exposure to activism, as Buolamwini goes public with her findings and undertakes an effort to create a movement toward accountability and transparency, including testifying before Congress to push for the first-ever legislation governing facial recognition in the United States and starting the Algorithmic Justice League.

These problems have most famously been explored by Safiya Noble (2018) in her book *Algorithms of Oppression*. Noble ultimately links these algorithmic problems back to capitalism, because they are created primarily by privately held companies whose main goal is to generate profit. Additionally, U.S. law of the past several decades has allowed many sites to function as monopolies that are able to purchase any potential competitors. A major example of this is Meta's purchases of Instagram and What's App. She explains this in greater deal in the following podcast:

Podcast: Algorithms of Oppression with Safiya Noble

Data & Society: Algorithms of Oppression

Episode Summary:

In “Algorithms of Oppression”, Safiya Umoja Noble challenges the idea that search engines like Google offer an equal playing field for all forms of ideas, identities, and activities. Data discrimination is a real social problem; Noble argues that the combination of private interests in promoting certain sites, along with the monopoly status of a relatively small number of Internet search engines, leads to a biased set of search algorithms that privilege whiteness and discriminate against people of color, specifically women of color.

Through an analysis of textual and media searches as well as extensive research on paid online advertising, Noble exposes a culture of racism and sexism in the way discoverability is created online. As search engines and their related companies grow in importance—operating as a source for email, a major vehicle for primary and secondary school learning, and beyond—understanding and reversing these disquieting trends and discriminatory practices is of utmost importance.

The capitalist imperative for profit is often either at the root of, or exacerbates these challenges. This is due in large part to the way that the internet has evolved and the way that many technology companies rely on advertising for their revenue. When a site relies on advertising to make money, they make more money the longer everyone stays on their site. This creates problematic outcomes, like YouTube's suggested viewing algorithm leading viewers to watch increasingly radicalized content (Sylvia and Moody, 2022). This approach has been dubbed the "Attention Economy," and you can learn more about its promises and perils in the following podcast:

Podcast: Adtech and the Attention Economy

Data & Society: Adtech and the Attention Economy

Episode Summary:

Data & Society Sociotechnical Security Researcher Moira Weigel hosts author Tim Hwang to discuss the way big tech financializes attention. Weigel and Hwang explore how the false promises of adtech are just one example of tech-solutionism's many fictions.

Of course, these problems are not limited to the United States, as they ripple out to the entire Global South. Racial capitalism is deeply ingrained in modern capitalist structures, affecting everything from labor markets to social movements. Exploring these challenges can be difficult. While racial capitalism was initially described as a form of data colonialism, recent scholars have suggested this may oversimplify what's happening. The podcast below, featuring Sareeta Marute and Emiliano Treré, explores the challenges while also highlighting possible avenues of resistance, underscoring the need for a critical examination of how data, race, and capitalism intersect in today's world.

Podcast: Data & Racial Capitalism

Data & Society: Data & Racial Capitalism

Episode Summary:

The conversation between the host and guests Sareeta Amrute and Emiliano Treré delves into complex issues such as digital activism, data colonialism, racial capitalism, and the Global South. Emiliano explores the challenges faced by indigenous and marginalized groups in Mexico, while both guests discuss the multifaceted nature of the Global South and critique the term "data colonialism." They also explore the pervasive algorithmic condition, the complexities of resistance, and the privilege and impossibility of disconnection. Sareeta's insights into IT workers in Berlin and their relationship with code highlight nuanced forms of resistance. The conversation concludes with an emphasis on everyday "counter conducts" and the importance of recognizing life outside of the algorithmic condition, offering hope for a more equitable and just future.

Additionally, it's important to consider feminist critiques of existing data practices. Data Feminism is an emerging field that intersects data science, feminism, and social justice, aiming to address the limitations of traditional data science methodologies. This approach applies an intersectional feminist lens to scrutinize who is involved in data collection, the purpose behind it, and the potential consequences for various communities. By doing so, it seeks to create a more ethical and inclusive data science practice that is sensitive to power dynamics, systemic inequalities, and context (D'Ignazio & Klein, 2020).

Ethical considerations are paramount in this interdisciplinary field, especially when dealing with big data collaborations between development organizations and large tech corporations. The concept of the "paradox of exposure" is introduced to question the benefits and risks of being counted in data sets, particularly for marginalized communities. This nuanced approach calls for participatory methods and co-creation to ensure that data collection and interpretation are both ethical and contextually appropriate (D'Ignazio & Klein, 2020).

The definition of what constitutes "data science" is also under scrutiny in this framework. Traditional definitions often marginalize interdisciplinary approaches and specific groups, particularly women and people of color. Data Feminism advocates for a broader, more inclusive definition that values ethical considerations and innovation from marginalized communities. This not only leads to more accurate and robust data science but also contributes to a more equitable and just society (D'Ignazio & Klein, 2020).

You can learn more about this in the following podcast, featuring the authors of the 2020 book, *Data Feminism:*

Podcast: Data Feminism

Data & Society: Data Feminism

Catherine D'Ignazio and Lauren F. Klein discuss their new book "Data Feminism," with Data & Society's Director of Research Sareeta Amrute.

Regulating Data

At this point, you may be wondering why we don't simply create better laws to address these issues with big data, and for example, prevent monopolies or the sale of social networks to foreign countries. While we could perhaps legislate the rules around how companies can be sold, regulating the actual use of big data turns out to be quite complicated. The reason for this goes back to the *why* question we addressed earlier, or rather the lack of the *why* question in the correlations made by big data. Let me explain.

Big data, by its nature, relies on the secondary usage of data, meaning it explores the connections between points of data that weren't understood or weren't the primary reason for collecting that data. An example of the primary use of data would be the collection of web-browser usage to understand how people are accessing a site and the most commonly used browser for which it should be best designed. A secondary usage of part of that data could be used to link browser usage to employment records in order to correlate browser choice with job performance. Browser usage data was not collected with that potential connection in mind, but a correlation was discovered in the data. Why is that true? My students love to speculate and try to create possible explanations, but the truth is, we simply don't know.

We could ban all secondary uses of data, but this would mean that we miss out on the good things big data can do: predicting outbreaks, preventing fires in New York City, fraud prevention, medical research on how wearables can predict upcoming heart attacks before they happen, etc. The point of big data is function creep. The function is the creep.

I've written elsewhere about potential regulation options that have been explored, but ultimately cannot be successful (Sylvia IV, 2016a). It's worth exploring these in detail to understand the significant challenges.

Notice and Consent

First, historically we have attempted to regulate data usage through notice and consent as part of the terms of service for a site or app. This approach is based on the 1980 Organization for Economic Co-operation and Development (OECD) Guidelines. The guidelines require users to be notified during sign up about what data will be collected and how it will be used. While this has always had limitations, it no longer even makes sense in the age of secondary uses of big data. Notice and consent is supposed to explain how your data will be used and give you the option to consent to that usage. While this is at least feasible for primary uses of data, we simply cannot know ahead of time what connections secondary uses of data will make. This means notice and consent practices have had to evolve to be so broad they essentially allow any use of the data generated, which more often than not passes through the servers of multiple different companies as part of analytics and ad serving processes. To truly understand how your data would be used, you would also need to read the notice and consent statement for every company through which your data passes.

The ability to read and understand such policies is also impacted by language barriers, especially for global technology companies. Many companies do not publish their terms of service or community guidelines in the languages of all of the people they serve. As of March 2019, Facebook translated their community standards into 41 of the 111 languages offered, Instagram 30 out of 51, WhatsApp 9 of 58, YouTube 40 of 80, Twitter 37 of 47, and Snapchat 13 of 21 (Fick and Dave, 2019). It's important to note users also encompass more languages than those officially supported by the platform. Additionally, Fick and Dave reported that Facebook translates the policies when a critical mass of users speak a specific language, but have no threshold for what they consider a critical mass.

There are additional challenges with this approach. Most sites have adopted a policy that allows only use or non-use of their site depending on whether or not you consent to the use of your data. If you don't consent, you don't get access to the services. The power dynamic here is tilted entirely in favor of the large corporations. If you're on the job market seeking a new position, how likely are you to opt out of using a service like LinkedIn if you don't fully agree with how they will use your data?

Further, these policies are difficult to read and time-consuming. A few years ago, I explored Facebook's terms of service only as they related to the use of data. An analysis showed that it would take the average person about 15 minutes to read that policy. Perhaps worse, the policy was written at an approximate average grade level of 13, meaning one would need at least some college education to be able to fully understand the policies. This is particularly problematic because 54% of adults in the U.S. are literate below the 6th-grade level (Rothwell, 2020). This puts white individuals, followed closely by Hispanics, at the greatest disadvantage because they have the highest rate of low literacy skills in the U.S. (35% White, 34% Hispanic, and 23% Black) (National Center for Educational Statistics, 2019). Researchers Lorrie Faith Cranor and Aleecia McDonald (2008) found the average length of a privacy policy to be 2,514 words, which would take the average person ten minutes to read. They then figured out that the average person visits between 1,354 and 1,518 websites in a given year. This comes out to requiring twenty-five full days a year, or seventy-six work days to read all of the policies associated with the websites we visit. Using some further calculations, they determine that if everyone in America read every privacy policy they're supposed to, it would add up to a nationalized total of 53.8 billion hours. This has likely increased quite significantly since this calculation was done in 2008.

We all joke about how no one reads these terms of service. But there's a reason. We couldn't possibly have enough time to actually read them. But most importantly, it's simply not possible to tell users what the secondary uses will be ahead of time.

Anonymization

One suggestion is built on the historically successful model of anonymizing data. However, it has become quite apparent that in the age of big data, the larger the data, and the more data sets that can be combined, the harder it becomes to truly anonymize any data in a way that prevents it from being anonymized by someone determined enough to do so. Many years ago, Chris Whong (2014) was able to access New York City taxicab data through a Freedom of Information Laws request. Although the data had been anonymized before being released, he was able to correlate data with publicly posted photographs to determine particular rides celebrities took, including how much (or how little!) they tipped. He was able to take this a step further by finding clusters of rides that dropped off in the same neighborhoods over time, and tie this to public records and social media accounts to identify a specific person who was regularly using taxis to visit gentlemen's clubs. This is a relatively straightforward example, but the larger point is that when enough data can be connected and correlated, deanonymization becomes much easier.

Deletion

Viktor Mayer-Schönberger (2009) has argued that we can make technical changes to how data is created and stored in computer systems. This proposed change would essentially allow all data to be given an automated deletion date. For example, all posts made to Twitter might be set to automatically delete after a one-year time period.

While this would certainly work from a technical standpoint, there are several practical challenges associated with this. For example, we would likely want to create the possibility to extend or change the date of deletion, which leaves open the possibility of such extensions happening indefinitely. This makes sense, as we may not want to automatically delete treasured family photographs, for instance. Furthermore, the question of who gets to set the deletion time period will be of utmost importance. If this is left to the corporations collecting data, they may simply extend the time period to be quite long.

Here, though, we have to also remember the deeper dynamics of big data. Even if we created new, incredibly strict regulations that put the power of choosing the time period for deletion into the hands of individual users rather than corporations, this approach would yet again risk losing some of the positive benefits that big data promises. For example, the heart rate data collected by wearables today might provide the data that an algorithm in 30 years time is able to use to predict and prevent the onset of various degenerative diseases. We might need significant longitudinal data to make exciting new correlational breakthroughs. These types of interventions would be most beneficial to the elderly and those with chronic diseases or cardiovascular risks (Chandrasekaran, 2020). Black adults and American Indians are twice and 1.5 times as likely to suffer from cardiovascular risks as White adults, so such advances could be especially helpful for those populations (Javed et al., 2022).

Regulate Harmful Uses

A Microsoft Global Privacy Summit (n.d.) suggested that regulators should focus on creating laws that prevent harmful uses of data. The discussions at this summit attempted to update the original OECD guidelines that promoted notice and consent. But these ultimately expanded the uses of data available to corporations so long as they weren't deemed harmful by "society," a deeply vague and problematic term. I further analyzed this proposal in this way:

> Rather than truly being guidelines for protecting the privacy of consumers, they are instead guidelines for managing the power wielded by corporations...
>
> Much of the data storage and processing is now done in the cloud, meaning through distributed computing. Big data projects are especially likely to be done this way because individual computers are often not powerful enough to process such large amounts of information, giving rise to services such as Apache's Hadoop, which offers just such distributed computing. This cloud computing, in combination with website services being distributed to so many third-party organizations, means that data flows are frequently crossing many different borders spanning organizations, nations, and most importantly, legal frameworks. Even if the United States were to create strong laws as a dissuasion to using data, it seems likely that data-reliant organizations would find a welcoming home in other countries with less strict laws. This process might, for instance, mirror those transformations in online gambling. Though illegal in the U.S., the servers are hosted in other countries, and still relatively accessible by U.S. citizens. (Sylvia IV, 2016)

Put simply, restrictive laws in one country might cause the servers to be moved to more lenient countries. In the case of online gambling cited above, there has been an increased push by several states to legalize and provide access to such gambling so that the taxes on such activities are not lost to other countries.

Ultimately, the biggest question here is who gets to decide what uses are harmful. The answer to that question moves out of the realm of privacy and into the realm of power and control.

AN OPEN QUESTION

Due to these challenges, it remains an open question how we might regulate the use of big data in ways that allow for its beneficial uses but prevent the harmful uses, at least in part because of challenges related to who gets to decide what counts as beneficial and harmful. Privacy protections would in theory allow users to decide when and what data of theirs to protect, but as we saw at the beginning of this chapter, privacy protections are in the midst of an erosion in the United States. Further, existing privacy protection only applies to materials located on one's own property, so any data that flows across the internet is not protected in that way.

I hope you can see clearly that these challenges related to big data and privacy apply to all of our daily lives. They are pressing, important, and difficult. But understanding what these challenges are is of utmost importance. The emergence of generative AI into prominence in 2022 and 2023 has made such questions even more pressing. Ethical discussion guides included in this book can be used to help start those conversations.

WRAP UP

Key Takeaways

- The issue of trust in technology companies is complex and varies across different demographic groups, with factors like race, gender, and educational level influencing how much personal data individuals are willing to share.
- Traditional methods for data regulation, such as notice and consent or anonymization, are becoming increasingly inadequate due to the complexities and secondary uses of big data, making it difficult to genuinely protect user privacy.
- The field of data science is grappling with ethical concerns, particularly around biases

that can affect marginalized communities; these biases are often unintentionally built into algorithms due to a lack of diversity among those who create and test technology.

- The regulation of big data faces significant challenges, including jurisdictional issues and the fundamental question of who gets to define what constitutes harmful or beneficial use of data, making it a complex issue of power and control.

Exercises

1. In what ways do you personally trust or distrust technology companies with your data? Do you think your race, gender, or educational level influences your level of trust? Discuss your reasons.

2. Choose one method of data regulation discussed in the material (e.g., notice and consent, anonymization, deletion) and argue its pros and cons. Can you suggest any modifications to make it more effective in the age of big data?

3. Listen to one of the podcasts mentioned in the material and summarize its key points. How does the podcast deepen your understanding of the ethical challenges posed by big data, and what solutions does it offer?

REFERENCES

Chandrasekaran, Ranganathan, Vipanchi Katthula, and Evangelos Moustakas. "Patterns of Use and Key Predictors for the Use of Wearable Health Care Devices by US Adults: Insights from a National Survey." *Journal of Medical Internet Research* 22, no. 10 (October 16, 2020): e22443. https://doi.org/10.2196/22443.

Conti-Cook, Cynthia. "Surveilling the Digital Abortion Diary." SSRN Scholarly Paper. Rochester, NY, October 28, 2020. https://doi.org/10.2139/ssrn.3666305.

D'Ignazio, Catherine, and Lauren F. Klein. *Data Feminism*. Strong Ideas Series. Cambridge, Massachusetts: The MIT Press, 2020.

Dobbs v. Jackson Women's Health Organization, 597 ___ U. S., 2022. https://supreme.justia.com/cases/federal/us/597/19-1392/case.pdf

DuBravac, Shawn. *Digital Destiny: How the New Age of Data Will Transform the Way We Work, Live, and Communicate*. Washington, DC: Regnery Publishing, a Salem Communications Company, 2015.

Greenwald, Glenn. *Why Privacy Matters*. TED Talks. TED, 2014. https://www.ted.com/talks/glenn_greenwald_why_privacy_matters?language=en.

Fick, Maggie, and Paresh Dave. "Facebook's Flood of Languages Leave It Struggling to Monitor Content." *Reuters*, April 23, 2019, sec. Media Industry. https://www.reuters.com/article/us-facebook-languages-insight-idUSKCN1RZ0DW.

Javed, Zulqarnain, Muhammad Haisum Maqsood, Tamer Yahya, Zahir Amin, Isaac Acquah, Javier Valero-Elizondo, Julia Andrieni, et al. "Race, Racism, and Cardiovascular Health: Applying a Social Determinants of Health Framework to Racial/Ethnic Disparities in Cardiovascular Disease." *Circulation: Cardiovascular Quality and Outcomes* 15, no. 1 (January 2022): e007917. https://doi.org/10.1161/CIRCOUTCOMES.121.007917.

Kates, Sean, Jonathan Ladd, and Joshua A. Tucker. "How Americans' Confidence in Technology Firms Has Dropped: Evidence from the Second Wave of the American Institutional Confidence Poll." Brookings, June 14, 2023. https://www.brookings.edu/articles/how-americans-confidence-in-technology-firms-has-dropped-evidence-from-the-second-wave-of-the-american-institutional-confidence-poll/.

Mayer-Schönberger, Viktor. *Delete the Virtue of Forgetting in the Digital Age*. Princeton, NJ: Princeton University Press, 2009.

McDonald, Aleecia M., and Lorie Faith Cranor. "The Cost of Reading Privacy Policies." *I/S: A Journal of Policy for the Information Society* 4, no. 3 (2008): 543–68.

Microsoft Global Privacy Summit Summary and Report Outcomes, pp. 1-24. Retrieved April 24, 2014, from http://www.vmsweb.net/attachments/pdf/NoticeandConsent.pdf

Moomaw, Graham. "Youngkin Administration Opposes Shielding Menstrual App Data from Search Warrants." NBC 12 News, February 15, 2023. https://www.nbc12.com/2023/02/15/youngkin-administration-opposes-shielding-menstrual-app-data-search-warrants/.

National Center for Educational Statistics. "Adult Literacy in the United States," July 2019. https://nces.ed.gov/pubs2019/2019179/index.asp.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press, 2018.

Oliver, John. *Net Neutrality: Last Week Tonight with John Oliver (HBO)*, (2014). https://www.youtube.com/watch?v=fpbOEoRrHyU.

Rothwell, Jonathan. "Assessing the Economic Gains of Eradicating Illiteracy Nationally and Regionally in the United States." Gallup, September 8, 2020. https://www.barbarabush.org/wp-content/uploads/2020/09/BBFoundation_GainsFromEradicatingIlliteracy_9_8.pdf.

Russonello, Giovanni. "How Big Tech Allows the Racial Wealth Gap to Persist." *The New York Times*, June 21, 2021, sec. U.S. https://www.nytimes.com/2021/06/21/us/politics/big-tech-racial-wealth-gap.html.

Sylvia IV, J.J. "The Ethical Implications of A/B and Multivariate E-Commerce Optimization Testing." In *Ethical Issues in E-Business: Models and Frameworks*. Hershey, NY: Business Science Reference, 2010.

Sylvia IV, J.J. "Little Brother: How Big Data Necessitates an Ethical Shift from Privacy to Power." In *Controversies in Digital Ethics*, edited by Amber Davisson and Paul Booth. Bloomsbury Academic, 2016a. https://doi.org/10.5040/9781501310553.

Sylvia IV, J.J. "Research." *J.J. Sylvia IV* (blog), August 29, 2016b. http://www.jjsylvia.com/research/.

Sylvia IV, J.J., and Kyle Moody. "False Information Narratives: The IRA's 2016 Presidential Election Facebook Campaign." In *Handbook of Research on Deception, Fake News, and Misinformation Online:*, edited by Innocent E. Chiluwa and Sergei A. Samoilenko. Advances in Media, Entertainment, and the Arts. IGI Global, 2019. https://doi.org/10.4018/978-1-5225-8535-0.

Sylvia IV, J.J., and Kyle Moody. "BreadTube Rising: How Modern Creators Use Cultural Formats to Spread Countercultural Ideology." CLCWeb: Comparative Literature and Culture 24, no. 1 (August 15, 2022). https://doi.org/10.7771/1481-4374.4291.

Whong, Chris. “FOILing NYC’s Taxi Trip Data” Chris Whong (blog), March 18, 2014. https://chriswhong.com/open-data/foil_nyc_taxi/.

PART II
GENERATIVE AI IN THE CLASSROOM AND WORKSPACE

Chapter Written by J.J. Sylvia IV and Elise Takehana[1]

Learning Objectives

- Explain the key differences between ChatGPT-3 and ChatGPT-4, including their capabilities and limitations.
- Develop an understanding of how generative AI can be utilized in various career paths, and

1. This guide was designed to be used in class as a way to introduce the topic of generative AI.

be able to critically assess the ethical and practical implications of its use in those fields.

- Acquire practical skills in generating effective prompts for ChatGPT, and will be able to evaluate the AI's outputs for quality, relevance, and potential biases.

GENERATIVE AI PRE-TEST

Please complete the Pre-Test in 15 minutes or less.

HOW GENERATIVE AI WORKS

Warm-up

Word Association

Choose a few of these to discuss as a group:

- The dog chased its [blank].
- I put my homework in my [blank].
- He hit the baseball with a [blank].
- She wore a beautiful red [blank].
- We watched the movie with a bucket of [blank].
- The teacher wrote on the [blank].
- During summer, I love to swim in the [blank].
- I read the entire book but didn’t understand the [blank].

- Every morning, she drinks a cup of [blank].
- He listened to his favorite song on the [blank].
- For my birthday, I got a new [blank].
- The astronaut looked out at the [blank].
- She likes to paint with water [blank].
- I play my favorite video game on the [blank].
- The athlete runs fast on the [blank].
- My favorite pizza is topped with [blank].
- The bear in the zoo loves to eat [blank].
- They cheered as their team scored a [blank].

There's also a connection here to cell phone predictive texting.

Figure 3: A screenshot of a cell phone text message being typed. The text being typed reads, "I put my homework in my," and the suggested autofills include, class, room, and car.

Markov chains use likelihood as predictors for the next in a sequence of words. How is this predicted? It's based on the texts that the model was trained on. Spend some time exploring these sources to better understand the process and the training data:

1. Jill Walker Rettberg – exploration of training data for GPT-3
2. What are large language models and how do they work?
3. How Generative AI Really Works
4. What is ChatGPT Doing and Why Does It Work?

DIFFERENCES BETWEEN CHATGPT 3 AND 4

1. GPT-4: How Is It Different from GPT-3.5?
    a. Amount of memory (25k to 3k).
    b. 40% more likely to generate factual information.
    c. More capable of reading emotions in the user.
    d. Performs better on standardized tests.
    e. Paid ChatGPT-4 now has live web browsing and other plugins, otherwise it can't browse the web. Need to change settings to access.

LINKS TO TOOLS

Spend some time exploring a few of the generative AI tools below:

- Open AI: ChatGPT: https://chat.openai.com/
- Google: Gemini: https://bard.google.com/
- Adobe: Firefly: https://firefly.adobe.com/
- DALL·E 2: https://labs.openai.com/

PROMPT-WRITING TIPS

How you write your prompts is a very important aspect of the quality of results that you get. Spend some time iterating the way you write prompts. We've also shared lots of links below that will help develop prompt writing skills.

1. Brainstorming Ideas / Articles:
    a. https://twitter.com/BrianRoemmele/status/1643032326652452864
    b. https://www.nytimes.com/2023/04/21/opinion/chatgpt-journalism.html
    c. https://www.nytimes.com/2023/05/25/technology/ai-chatbot-chatgpt-prompts.html
    d. https://www.oneusefulthing.org/p/a-guide-to-prompting-ai-for-what
    e. https://www.oneusefulthing.org/p/how-to-use-ai-to-do-practical-stuff
    f. https://prompts.chat/
    g. https://twitter.com/MushtaqBilalPhD/status/1621379333943083009
    h. https://twitter.com/MushtaqBilalPhD/status/1637715972705468417
    i. https://twitter.com/thatroblennon/status/

1610316022174683136

j. For Images:
   i. https://artificialcorner.com/youre-using-midjourney-wrong-here-s-how-to-create-better-images-than-99-of-midjourney-users-c876fbe7915e
   ii. https://letsenhance.io/blog/article/ai-text-prompt-guide/

2. Suggested Tips:
   a. It performs better when you provide it info rather than ask it for info. This is a way to avoid hallucination problems. Try "Summarize the following text:" or "Explain the following text at an 8th grade reading level:" Consider using some complex academic article abstracts to test this.
      i. GPT3.5 cannot follow a URL to get information, but it will hallucinate content and appear that it can do so if the URL is descriptive enough. You can't simply paste in a URL, you should paste the content directly in.
   b. Ask it to re-write text in different styles. "Rewrite this in the style of..."
   c. It tends to perform better when you assign it a role and give it a task and format. "Act as... to complete the task of ... and maintain the format of..." Examples here.

        i. Act as a professor and write...
            1. Acting as a professor, your task is to design a syllabus for a class on data and society. Include a weekly reading list. Please use only real verifiable sources that are cited. The class should be appropriate for college freshmen.
        ii. Act as a life coach...
        iii. Act as a senior front-end developer
        iv. Act as a nurse...
        v. Act as a travel agent...
    d. Ask it to generate text-to-image for other AI tools like Midjournay/DALL-E 2
    e. Create an interactive choose-your-own adventure game.
        1. "Create an interactive choose-your-own adventure game about Star Trek. I will play the role of the captain. You will prompt me with multiple choice options for what actions I will take, but also allow me to give my own answers that go beyond the choices."
    f. Use a temperature setting between 0.1 and 1. 1 is more creative, more likely to hallucinate, more unpredictable. 0.1 is the most stable and confident result.
3. Iterative prompting
    a. We recently drafted a chapter for an edited

collection about incorporating AI, and included a supplemental page with iterative prompting examples: http://www.jjsylvia.com/wicked-ai/

b. If you have a very long prompt generation, you may need to type "continue" to have it finish the text generation.

AI CAREER RESEARCH

Objectives:

1. Develop an understanding of the diverse ways AI is likely to be utilized in future workplaces.
2. Research a specific career path of interest and its potential interaction with AI.
3. Share your findings.

Instructions:

Part 1: Initial Research

1. Choose a specific career path you are interested in.
2. Conduct initial research on the current status of AI in that field. Some potential questions to guide your research might include:
    - How is AI currently being utilized in this field?
    - What specific tasks or roles are being automated or assisted by AI?
    - What are the benefits and potential drawbacks of

this AI integration?

Part 2: Future Forecasting

Based on your research and understanding of AI capabilities, predict how AI might further influence this career path in the next 10-20 years. Consider the following aspects:

1. What additional tasks or roles could be automated or assisted by AI?
2. What new opportunities might arise due to AI integration?
3. What challenges could professionals in this field face due to increased AI use?

Part 3: Presentation Creation

We're going to create a collaborative presentation. Open the following Google Slides deck and then add one slide. Share your findings with text and images.

Articles We Found Across the Disciplines

Libraries: Hennig, Nicole, and Daniel Pfeiffer. "A Tech Librarian Explains How to Build AI Literacy," April 24,

2023. https://www.choice360.org/libtech-insight/a-tech-librarian-explains-how-to-build-ai-literacy/.

Computer Science: Cheng, Michelle. “Coding Instructors Are Adding AI to Their Lessons—before AI Replaces Them,” May 8, 2023. https://qz.com/coding-instructors-are-adding-ai-to-their-lessons-befor-1850393865.

Video: *Utilizing AI for Documentary Production – with Basil Shadid and Philip Shane*, 2023**.** https://www.youtube.com/watch?v=9dP3mYZ4FR8.

Nursing: University of Calgary. “From curiosity to care: A mindful integration of AI in nursing education,” May 8, 2023. https://nursing.ucalgary.ca/news/curiosity-care-mindful-integration-ai-nursing-education

Sociology: Balmer, Andrew. “A Sociological Conversation with ChatGPT about AI Ethics, Affect and Reflexivity,” May 3, 2023. https://journals.sagepub.com/doi/full/10.1177/00380385231169676

Speech: Haynes, James. **“**What ChatGPT and AI can do for speakers,**”** https://thespeakerlab.com/what-chatgpt-and-ai-can-do-for-speakers/

Psychology: Ruiz, Rebecca. “3 things to know before talking to ChatGPT about your mental health” KJanuary 30, 2023, https://mashable.com/article/how-to-chat-with-chatgpt-mental-health-therapy

Education: Heaven, Will Douglas “ChatGPT is going to change education, not destroy it.” April 6, 2023.

https://www.technologyreview.com/2023/04/06/1071059/chatgpt-change-not-destroy-education-openai/

Business: Matt Symonds. “More Diversity and Opportunity, Less Trigonometry – The Future of Graduate Management Education and the GMAT.” April 27, 2023. https://www.forbes.com/sites/mattsymonds/2023/04/27/more-diversity-and-opportunity-less-trigonometrythe-future-of-graduate-management-education-and-the-gmat/?sh=22ca76914438

Mathematics: Ferlazzo, Larry. “How Teachers Are Using Artificial Intelligence in Classes Today,” May 2, 2023. https://www.edweek.org/technology/opinion-how-teachers-are-using-artificial-intelligence-in-classes-today/2023/05

Journalism: Manjoo, Farhad. “ChatGPT Is Already Changing How I Do My Job,” NY Times. April 21, 2023 https://www.nytimes.com/2023/04/21/opinion/chatgpt-journalism.html

Journalism: Carlson, Nicholas. “My editor’s note to the newsroom on AI: Let’s think of it like a ‘bicycle of the mind,” Business Insider. April 13, 2023 https://www.businessinsider.com/how-insider-newsroom-will-use-ai-2023-4

HANDS-ON PROJECT

Social Media Campaign

For this activity, you're going to create a social media campaign: Social Media Assignment

Essay Reflection

Find an essay or other piece of writing that you've previously created for an assignment in school. Share it with ChatGPT and then ask ChatGPT to comment on the quality of the writing of your essay and paste the essay in the prompt. Think about how much you agree or disagree with its comments and ask several follow-up questions that ask for specific examples from your essay. Try asking it to make suggestions about sources to use or ways to rewrite your work, but remember that ChatGPT isn't there to "fix" your writing. Think critically about what ChatGPT values in writing and what you value.

Example: https://chat.openai.com/share/651420bc-8662-4623-aaa1-02610eeeac63

Accompanying Questions

- Comment on the quality of the writing of this student essay: PASTED THE ESSAY HERE.
- Can you provide specific examples from the student essay where the style was weak and the points were not well supported by evidence?
- Why do you think the student's introduction and conclusion were weak? Provide examples from the essay in your response.
- Can you rewrite some of the most boring sentences in the essay in a more engaging way?
- Can you recommend some sources that this student could use to help them develop their essay? Include links when available.
- What are some counterarguments or oversights a critical reader could have of this student's essay's position?
- What are the strongest points the essay makes?

DISCUSSION OR REFLECTION QUESTIONS

Article for discussion kick-off:

Here are the top skills you will need for an 'A.I.-powered future,' according to new Microsoft data

Discussion:

1. What are the potential advantages and disadvantages of using AI tools, such as generative AI, to assist with school work?
2. If a student uses an AI tool to write an essay or complete a project, who should receive credit for the work – the student, the AI, or both? Why?
3. Can the use of AI tools for academic tasks be considered a form of cheating? Why or why not?
4. How might the use of AI tools for academic work influence a student's learning process and development of critical thinking skills?

5. In what ways might the use of AI tools for academic work affect the teacher-student relationship and academic evaluation processes?
6. Do students have a responsibility to disclose when they've used AI tools for school work? Why or why not?
7. What guidelines or policies could schools implement to govern the use of AI tools in academic work?
8. How does the use of AI tools in school work raise questions about the nature and purpose of education?
9. How might socioeconomic disparities in access to AI tools affect academic fairness and equity?
10. How can the educational sector ensure that the use of AI tools aligns with academic integrity principles and promotes equitable educational outcomes?
11. How do you think AI should be used in the classroom? For assignments? What would you want your teachers and future professors to know about AI?

LANGUAGE, DIVERSITY, INCLUSIVITY, AND CHAPTGPT

Guiding Questions

1. Knowing what ChatGPT is trained on (search engine crawl, ebooks, reddit, and wikipedia), what kinds of cultural concepts or groups might not be included?
    - What about oral languages, since less than 10% of human languages are written?
    - What about non-standard inscription media like the Benin bronzes, Incan quipu, or Maori carvings?
    - Is a translation ever an accurate representation of the original?
2. What languages do you speak and what have you noticed about moving back and forth from those languages?
    - "How To Speak Bad English" (8:20-13:40) podcast episode on Global English and accent reduction. A major point here is that more English speakers are

"nonnative" than "native" so "native" speakers need to adjust their expectations on what "clear communication" is.

- Visualizing the Most Used Languages on the Internet

Discussion Questions:

For the group discussing "ChatGPT threatens language diversity":

1. How does the AI respond to prompts in non-English languages?
2. Does the AI show any bias towards English language or syntax when generating responses?
3. Try typing a sentence with non-English syntax in English. How does the AI respond?

For the group discussing "ChatGPT is multilingual but monocultural":

1. Generate a story set in a non-Western culture. Does the AI accurately and respectfully incorporate elements of

that culture?
2. How does the AI respond to prompts containing cultural idioms, references, or concepts?
3. Look up some common phrases or idioms in less commonly used languages. How does the AI respond to these prompts?

For the group discussing "Proper English and normative grading practices":

1. Try typing sentences in various English dialects or accents (e.g., African American Vernacular English, Singlish, Hinglish). How does the AI respond?
2. Does the AI seem to favor a particular type of English in its responses?
3. How does ChatGPT's answer to your question change as you rephrase the same question (ie using "Black" rather than "African American") Does it perpetuate stereotypes or exhibit biases?

Links to Useful articles on this

Summary points

a. "Unmasking AI Harms and Biases".
b. "ChatGPT threatens language diversity in the age of AI"
    i. With white male voices authoring the majority of the training material, the default voice replicates those language patterns.
c. "ChatGPT is multilingual but monocultural, and it's learning your values"
    i. Diversity is not just in languages and dialects used but in the cultural beliefs and ideologies embedded in the training material.
d. "ChatGPT & Writing in the Secondary ELA Classroom"
    i. Our normative grading practices around "proper English" encourage students to mask language diversity.
e. "OpenAI's Linguistic Diversity Initiatives in AI Language Testing"
    i. OpenAI's proposed solutions focus more on including less common languages but say much less about how to addressing race and gender stereotypes in language use.

POST-TEST AND SURVEY

Please complete the following Post-Test and Survey in 30 minutes or less.

An interactive H5P element has been excluded from this version of the text. You can view it online here: https://rotel.pressbooks.pub/datarenaissance/?p=74#h5p-3

WRAP UP

Key Takeaways

- Generative AI models like ChatGPT-4 have evolved significantly in their capabilities, including better factual accuracy and emotional understanding, but they come with limitations such as the inability to browse the web unless specific settings are enabled.
- The use of AI tools in academic and professional settings raises important ethical questions around authorship, academic integrity, and the potential for perpetuating biases or excluding certain cultural perspectives.
- Prompt-writing techniques greatly influence the quality of outputs from generative AI;

understanding how to effectively structure prompts can lead to more accurate and useful responses.

- The role of AI in various career paths is not just an imminent future but a present reality, necessitating research and ethical considerations about how AI will shape and be shaped by professional practices.

Exercises

1. How do you think the use of AI tools like ChatGPT could affect language diversity and inclusivity? Consider ChatGPT's training data and its implications for representing various cultures and languages.
2. Using the generative AI tools listed in the "Links to Tools" section, each student is

tasked with creating a piece of content (it could be text, art, or any form of digital media). Afterward, discuss as a class the ethical considerations you had to make while using these tools. Were you concerned about the originality of your work, the biases in the AI, or other ethical issues?

3. Write a prompt for ChatGPT that aims to generate a summary of a complex academic article. Evaluate the accuracy and clarity of the AI-generated summary. What does this exercise reveal about the strengths and limitations of using AI for academic purposes?

FURTHER READING

"AI Text Generators: Sources to Stimulate Discussion among Teachers." https://docs.google.com/document/u/0/d/1V1drRG1XlWTBrEwgGqd-cCySUB12JrcoamB5i16-Ezw/mobilebasic.https://www.theatlantic.com/technology/archive/2023/02/chatgpt-ai-detector-machine-learning-technology-bureaucracy/672927/

Fyfe, Paul. "How to Cheat on Your Final Paper: Assigning AI for Student Writing." *AI & SOCIETY*, March 10, 2022. https://doi.org/10.1007/s00146-022-01397-z.

Bogost, Ian. "ChatGPT Is About to Dump More Work on Everyone." The Atlantic, February 2, 2023. https://www.theatlantic.com/technology/archive/2023/02/chatgpt-ai-detector-machine-learning-technology-bureaucracy/672927/.

Hoffman, Reid, and ChatGPT. *Impromptu*. 2023. https://www.impromptubook.com/wp-content/uploads/2023/03/impromptu-rh.pdf.

Knight, Will. "These ChatGPT Rivals Are Designed to Play with Your Emotions." May 4, 2023. https://www.wired.com/story/fast-forward-chatgpt-rivals-emotions/

Ray, Augie. "What Does It Mean If AI Can Convey More Empathy Than Humans?," 2023. https://www.linkedin.com/pulse/what-does-mean-ai-can-convey-more-empathy-than-humans-augie-ray.

Sentient Syllabus Project. "Syllabus Resources." Google Docs, 2023. https://docs.google.com/document/d/1O1_uUvF8OYbleru5QyjjuNP_On7h5vaVQC2GaSQ315U/edit?usp=sharing&usp=embed_facebook.

"Sparks of Artificial General Intelligence: Early Experiments with GPT-4," https://arxiv.org/pdf/2303.12712.pdf

The New York Times. "Hard Fork." https://www.nytimes.com/column/hard-fork.

UNESCO. "ChatGPT and Artificial Intelligence in Higher Education: Quick Start Guide," 2023. https://www.iesalc.unesco.org/wp-content/uploads/2023/04/ChatGPT-and-Artificial-Intelligence-in-higher-education-Quick-Start-guide_EN_FINAL.pdf.

Wong, Matteo. "ChatGPT is Already Obsolete," May 19, 2023. https://www.theatlantic.com/technology/archive/2023/05/ai-advancements-multimodal-models/674113/.

PART III

CASE STUDY: "IT'S PERFECT, FOUR STARS!"

Chapter Written by Leonora Shell, M.A.T.[1]

Editor's Note: This chapter is written from a first-person perspective by a woman who owns a business that sells products on Etsy. It is intended to highlight her personal experience as a woman.

Learning Objectives

1. This chapter was written based on a guest lecture that was given to both the undergraduate and ALFA sections of the Data & Society course. In the spirit of full disclosure, Leonora Shell is the wife of volume editor J.J. Sylvia IV. To avoid conflict of interest, the shop referenced has not been disclosed, either here or in the guest lectures.

- Critically analyze the impact of review systems on e-commerce platforms, particularly how they affect small business owners and vulnerable populations.
- Gain an understanding of how algorithms and automated systems can both support and hinder the operation of online marketplaces, influencing the livelihoods of the sellers involved.
- Develop the ability to evaluate the ethical implications of review and rating systems in digital commerce.
- Analyze the effects of review systems on power dynamics and societal inequalities through a feminist lens.

INTRODUCTION

Black Mirror's 2016 episode, "Nosedive" explores a dystopian future in which the protagonist's life is ruined as she accidentally lowers her overall personal rating over the course of a very bad day. The idea of a personal rating is already starting to take shape in the form of China's social credit system, but even in the U.S., we have an analogous system of rating and ranking, even if it hasn't been centralized in one consolidated place.

The effects of starred ratings and reviews for consumer products have been heralded as a way to create an objective, quantifiable method for assessing the quality of a product or service (Gunasekaran, 2019). On the surface, this seems to be true, a way to summarize a consumer experience using a simple five starred approach, ranging from five stars meaning you loved it, to one star being "disappointed." More often than not, however, these ratings are not about the particular good or service, but more about the mismanagement of expectations by the consumer (Peak Performance Digital, n.d.). Furthermore, negative ratings are often unaccompanied by any sort of relevant commentary or a way for a company or individual seller to improve. As more women enter the space of e-commerce and business, the reviews have taken on more sexist and harmful tones as well as the introduction of AI or automated bots that crawl sites and take down a seller's listings without warning or an effective way to counter the decision that didn't involve a human's judgment at all.

HUMAN COMMERCE

After over ten years of selling handmade products on an e-commerce site specifically designed for handmade goods, one that touts the importance of keeping commerce human, while continually and methodically removing any empathy from reviews or oversight from actual human beings, it is clear that a linchpin moment in the change to a less qualitative and human experience in the marketplace was the transition to starred reviews by customers in late 2013. With this, there was also the removal of the option to rate or review customers by a seller.

As a business owner writing this overview of the impact of starred reviews, I wanted to share my insights as an individual shop being rated on an ever-changing platform in an economy that not only demands constant growth but also perfection. I am a professional, and negative reviews are generally a place to learn and grow; they certainly don't bother me like they used to — as of this writing I have over 23,000 sales, over 3,300 reviews and over 10,000 followers on social media, a community built organically over a decade around our brand. As a result of our success as a brand, we have never had to take on outside funding, loans or debts. We have been extremely lucky and grateful for the tools that we have used to get us where we are today. So much of what we do, sell, and market has success or failure based on algorithms created by other publicly-traded companies. Since this is my current full-time employment, I will not be using the name of the company in this chapter, as it is against their current user policies to do so in a negative (albeit truthful) way.

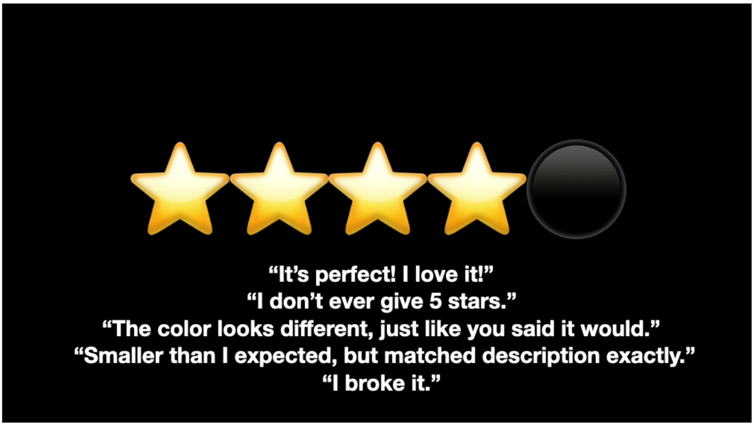
LIVING AND DYING BY THE ALGORITHMS

So, how does the algorithm work? This is the great question with e-commerce and online marketing across various social media, search, and commerce platforms. Since the handmade e-commerce site in question went public in 2015, the priorities of the company change after each quarterly board meeting. There is a cycle and culture of pushing for ways to endlessly improve and be tinkered with — AI and bots crawl the site and can remove and shut down shops with little to no warning. Trends can be decided upon and implemented based on keyword searches of a couple hundred individuals, the lists of celebrity influencers, a feature in an advertisement or a selection by an employee of the company.

We are here to focus on the reviews, however, as no marketplace that exists is perfect for consumers and business owners. In essence, the higher a given shop's reviews, the more frequently their products appear in search results, thus resulting in more sales (Collinger and Malthouse, 2015). This is good, right? Because as a consumer, you would want only the best products shown to you when you search for them rather than the worst or least popular — but what if that popularity was artificially downgraded for smaller sellers and falsely upgraded for drop-shippers or sellers that weren't hand-making their products? The site attempted to remedy any customer dissatisfaction with a mathematical formula, the details of which were hidden from both sellers and buyers, called the the Order Dissatisfaction Rating (ODR) (Glassenberg, 2020). This ODR included the amount of customer complaint cases brought against a seller and shop ratings, among other metrics. If your ODR rose to an amount higher than was defined as acceptable (at the time less than 1% of reviews could be 1 or 2 stars, over the course of 90 days), your shop would be warned, closed and/or you would be suspended from using the site. Unfortunately, only about 10% of purchasers review their orders, and this metric only took into account the reviews that were made, not the overall percentage of orders fulfilled with satisfied customers.

The ODR approach was ended in 2020 because so many shop dissatisfaction ratings increased after customers were unhappy with the disruptions to shipping and the overall logistics of the planet's supply chain due to the COVID-19 pandemic. Those customers that felt out of control took it out on sellers in the form of low reviews. In 2021, a program was created called the "Star Seller Program," which was a way to display the ODR to consumers and sellers. Previously, this ODR metric was only accessed with a non-navigable link which was made available on a Reddit post or discussion boards hosted by other organizations under a now-deleted section of the site called "customer service performance." The link is now broken. This new program began in 2021 and required that 95% of a shop's reviews over 90 days be five star reviews, meaning that for every 4 star review a shop needed 19 five star reviews to maintain their standing as a star seller. If a shop only receives reviews on 10% of their sales, a single 4-star review would require 190 additional sales with exclusively five star reviews to recover their star seller standing. In 2022, after considerable seller feedback, this policy was remedied so that a shop had to maintain a 4.8 average star rating over 90 days, significantly increasing the ability for shops to be a part of the program that sets them apart from other sellers.

THE “FAULT” IN OUR STARS



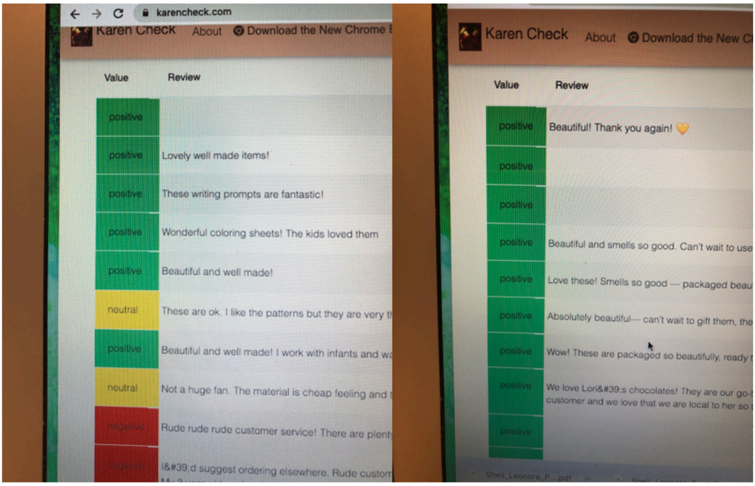
Some favorite four star reviews over the last ten years.

Figure 1. The image features four stars and several quotes from user reviews. These read: ”It’s perfect! I love it!”, “I don’t ever give 5 stars.”, The color looks different, just like you said it would.”, “Smaller than expected, but matched description exactly.”, and “I broke it.”

Let’s take a brief aside to evaluate four star ratings. Some of my most frustrating reviews over the past decade have been four star reviews. After a brief analysis of these four star reviews, almost half of those our shop received contained the words “perfect” and “love.” When kindly questioned if there was anything we could have done to help us get a perfect score, several customers stated that they simply don’t give five star reviews to any products ever. I have stopped asking this question as some consumers retaliate and end up lowering their initial review, citing that all we cared about was a review. Other customers state that the color is different than pictured, just like it was clearly stated in the listing and/or as the customer requested a custom color or variation… and so it did, indeed, vary. Maybe the item was smaller than they expected, however it matched the description exactly, and they loved it. Or, more maddeningly, they broke it and didn’t let us know or refused to take a replacement despite our robust policies stating that if they break an item we will replace it, no questions asked. There is no policy, kindness, or gesture that can remedy these situations, and that’s just how business works sometimes.

WHO RATES THE RATER?

Historically, sellers on the handmade e-commerce site were able to leave tiered reviews for customers that were negative, neutral or positive; however, with the establishment of starred reviews, this ability was removed by the site and later replaced with only the ability to block buyers rather than provide any feedback. A site that was helpful when making decisions about how to engage with a customer was called “KarenCheck.com” — this site has now been hobbled, as it has been blocked from accessing reviews connected with usernames on the e-commerce site in question. However, for a time it was useful to see the history and types of reviews left by a customer to check their expectations and ways of communication they were most likely to positively respond.



Screenshots of KarenCheck.com before it was effectively banned by the marketplace.

Figure 2. Screenshots of KarenCheck.com before it was effectively banned by the marketplace. The image shows one reviewer with a mix of positive, neutral, and negative reviews, and another reviewer with all positive reviews.

Additionally, in 2021, social media sites such as TikTok encouraged customers to scam sellers to get free items by leaving negative reviews and demanding returns, only to return trash from their homes or something else to the seller. Our small shop personally received a box of hair and another customer used glittered paper to fill their returned package, causing us to have to do an additional round of completely sterilizing and cleaning our workshop.

THE SHIFT TO "OBJECTIVE" STARS

So, where did all of this start? In early 2013, the reviews that customers could leave featured three options: "negative," "neutral," or "positive." In late 2013, these values were switched to a starred approach, with negative reviews being translated to one or two stars, one star meaning "disappointed" and two stars meaning "Not a fan." Neutral translated to three and four star ratings meaning "It's okay." and "Like it," respectively. Lastly, positive reviews were translated to five stars, meaning "love it." These words pop up when a user hovers over the number of stars to select which one best describes their shopping experience. The resulting "experiences" for the customer shifted from a qualitative, subjective, generic feedback of experience approach for leaving product reviews to a quantitative, objective, specific and rating/grade approach. The shift from seller or overall shop review to the review of an individual item, which varies, and by the nature of the marketplace, should be handmade and individually created for each customer. This method is not only dehumanizing for the seller, but gives the buyer a significant amount of leverage in the future of a shop or item in a given marketplace because of the way it impacts the algorithms that determine which items are featured in searches.

Slide from presentation given by the author on this topic.

Figure 3. Slide from presentation given by author. It features the differences in the rating system from early 2013 to late 2013, noting changes from qualitative to quantitative, subjective to object, generic to specific, and feedback/experience to rating/grade.

All of this in itself could be considered fine; it is a way to create a seemingly equitable marketplace for sellers with a transparent review system allowing freedom of expression of contentment or discontentment with a particular item or service. This could absolutely be said about a marketplace wherein each seller and buyer represent the same demographics, but not one where the sellers and buyers represent populations with historically very different societal power and autonomy.

WHO ARE WE RATING?

So, who are we rating anyway? Why do we care about stars versus negative/neutral/positive experience metrics? Of course, we need some way to assess sellers and their products, but I argue that this metric unequally impacts the most vulnerable business owners. As of the most recent report of this U.S. marketplace at the time of this writing, 86% of sellers on this platform identified as female (Drah, 2021). These are employment opportunities for women to be business and micro-business owners that are self-made. They are not involved in multi-level marketing schemes or employed by someone else; these are ways women can make their own hours and own their own businesses. For many years, this platform encouraged its users to quit their day jobs and go full time with their craft. These sellers are also twice as likely to be under the age of 35, with a median age of 39. This represents a lot of work-at-home parents, those between more stable employment, or part-time workers. Aligned with the national income average, 17% of sellers make less than $25K per household. Additionally, 97% of these shops are run from home. On the other side of things, 70% of buyers identify as female. However, the site has announced that they are encouraging and focusing on bringing more men to the platform in 2023 (Ryan, 2022).

According to Pew research (Smith and Anderson, 2016), men are more likely to read and leave reviews and younger consumers are more likely to leave reviews. Anecdotally, those men have been more critical, citing that they never leave five star reviews, are unwilling to change a review based on new information, or even accept a refund. This gives me pause, as this means that the site is bringing a potentially more critical population to rate and review predominantly young women, who are historically underemployed and underrepresented in business ownership opportunities and spaces (Lake, 2023). And, despite women nearly reaching gender parity in 2023 in business education programs, they still aren't compensated or funded equally — in most cases making half of their male counterparts (Arora, 2020).

CONCLUSION

So, what are we to do in a culture driven by capitalism? According to Cory Doctorow (2023), the concept of “enshittification” explains the market forces that encourage a platform to cater to different strata of the population over time. The platform is successful in keeping each group happy and dependent on the platform until they shift methods and gears to bring in another group in order to grow, ultimately frustrating everyone who was already there, leaving those initial users who were early adopters of the platform and brought it to existence and initial success, in the dust. This enshittification applies to the handmade marketplace in question, leaving its original handmade sellers and early devotees in the throes of quirky algorithms and shifting priorities after almost each quarterly shareholder meeting. This cautionary tale leaves everyone to consider, and reconsider, their place in a business world where various platforms vie for their and others’ time, attention, and most importantly, money. Each year, an additional small percentage is added to the amount of each sale that is claimed by the platform, with less and less value-add. This extra cost burden gets passed on to the consumer, or more likely, the seller to stay competitive.

Now, if I get a negative review, I strive to respond in a more productive way. I leave good reviews for places I love – small businesses that are making it work despite so many pressures to close up shop and go work for someone else, so many pressures to abandon their own good ideas and lives and devote them instead to making more millionaires become billionaires. Five stars go to the small restaurants in our neighborhood that I want to keep open. When people feel out of control of their lives, they receive bad news, or something sad and out of their control happens, they attempt to take some control in some way, and sometimes that's leaving a negative review. In my case, it's to leave a positive one. It's a very human thing to do, it turns out. A way to keep commerce human.

I remind myself that this one e-commerce platform is just that: it's just one place that I choose to spend my time, ideas, energy, and money. Over the last decade, I've created several supportive environments outside of the e-commerce site that allow connection with customers and positive experiences with families using our handmade products. I won't let the design of the platform or the whims of the shareholders detract from my positive experience. To do that is to lose what remaining joy I have and bring to what I make and share with others. And no bad review can take that away from me. A review is not my identity, it is not me. I am the only one who can manage my own expectations and I encourage you to do the same the next time you are asked to leave a review.

WRAP UP

Key Takeaways

- The shift from qualitative to quantitative review systems in e-commerce platforms has had a profound impact on small business owners, often amplifying the power imbalance between sellers and consumers.
- Algorithms and automated systems, while designed to improve marketplace efficiency, can inadvertently penalize sellers through non-transparent metrics and sudden policy changes.
- The demographic makeup of a platform's user base can influence the nature and tone of reviews, with evidence suggesting that

women and younger business owners might be disproportionately affected.

- The concept of “enshittification” encapsulates the risk that platforms, in their quest for growth, can alienate their original user base, undermining the very communities that contributed to their initial success.

Exercises

1. Consider the ethical implications of relying solely on algorithms to manage reviews and seller standings in an e-commerce platform. How can these systems be improved to account for the human element in commerce?
2. Explore the concept of “enshittification” in the context of other online platforms or services you have used. Can you identify any instances

where a platform's changes have alienated its original user base? Discuss the long-term sustainability of such strategies.

3. Given the gender and age demographics outlined in the text, analyze how these might interact with the review and rating systems to create a potentially biased marketplace. What steps could platforms take to mitigate these biases?

REFERENCES

Arora, Rohit. "Why Male Entrepreneurs in the US Make Double Their Female Counterparts." CNBC, March 10, 2020. https://www.cnbc.com/2020/03/10/why-male-entrepreneurs-in-the-us-make-double-their-female-counterparts.html.

Collinger, Tom, and Edward C. Malthouse. "From Reviews to Revenue Volume 1: How Star Ratings and Review Content Influence Purchase." PowerReviews. Northwestern University, 2015. https://www.powerreviews.com/wp-content/uploads/2019/02/From-Reviews-to-Revenue-Northwestern-Report-Volume-1.pdf.

Doctorow, Cory. "Pluralistic: Tiktok's Enshittification (21 Jan 2023) – Pluralistic: Daily Links from Cory Doctorow." *Pluralistic: Daily Links from Cory Doctorow* (blog), February 27, 2023. https://pluralistic.net/2023/01/21/potemkin-ai/.

Drah, Hermina. "19 Etsy Statistics for a Profitable 2023." *Why Does Everything Suck?* (blog), June 10, 2021. http://whydoeseverythingsuck.net/blog/etsy-statistics/.

Glassenberg, Abby. "ODR on Etsy: What Sellers Need to Know." Craft Industry Alliance, February 18, 2020.

https://craftindustryalliance.org/odr-on-etsy-what-sellers-need-to-know/.

Gunasekaran, Ajay. "Why Star Ratings Don't Make Sense for All Product & Services." Medium, 2019. https://uxdesign.cc/why-star-ratings-dont-make-sense-for-all-product-services-908e3b3901db.

Lake, Sydney. "MBA Programs Are Nearly Reaching Gender Parity with More than 41% Women Enrollment." Fortune, January 27, 2023. https://fortune.com/education/articles/mba-programs-are-nearly-reaching-gender-parity-with-more-than-41-women-enrollment/.

"Nosedive." Streaming. *Black Mirror*. Netflix, 2016.

Peak Performance Digital. "5 Stars or Bust: Why the Star Rating System for Reviews Is Broken – Peak Performance Digital," n.d. https://peakperformancedigital.com/5-stars-or-bust-why-the-star-rating-system-for-reviews-is-broken/.

Ryan, Tom. "Has Etsy Been Missing the Boat with Men?" *RetailWire* (blog), May 9, 2022. https://retailwire.com/discussion/has-etsy-been-missing-the-boat-with-men/.

Smith, Aaron, and Monica Anderson. "Online Shopping and E-Commerce: 2. Online Reviews." *Pew Research Center* (blog), December 19, 2016. https://www.pewresearch.org/internet/2016/12/19/online-reviews/.

PART IV
MEDIA AND DATA LITERACY

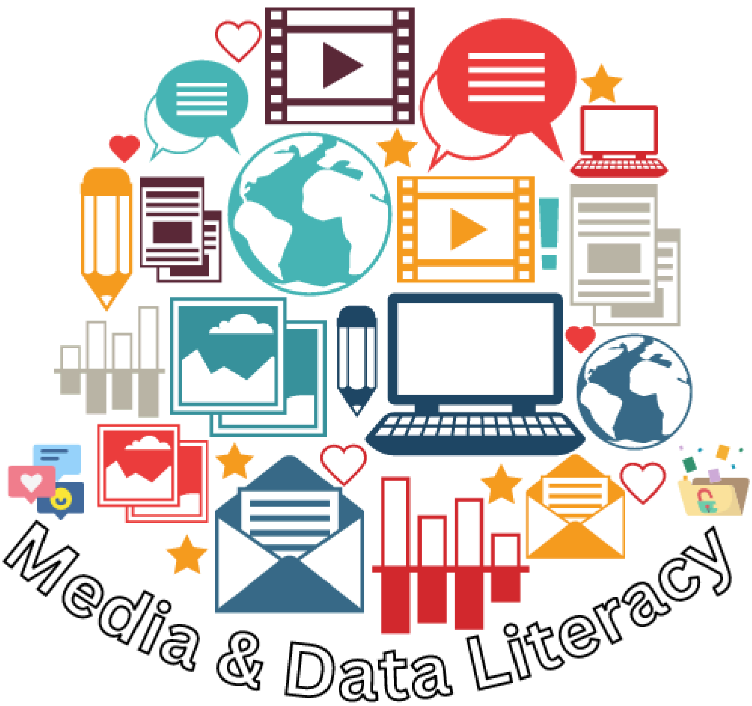


Figure 1. Media & Data Literacy. Media and data is being shown with pencils and paper.

Chapter Written by Henry Christiansen

Learning Objectives

- Critically evaluate various forms of media and data, identifying biases and ethical implications inherent in them.
- Gain a comprehensive understanding of both media literacy and data literacy, and how these two fields intersect in various aspects like race, gender, and social class.
- Understand the challenges and opportunities in media and data literacy education, and be equipped with strategies for incorporating these literacies into different learning environments and career paths.

INTRODUCTION

Currently, media literacy and data literacy are still relatively new and evolving fields. It's important to acknowledge that access to these forms of literacy is not uniform across all communities. Systemic barriers, such as socio-economic status, educational opportunities, and cultural factors, often create a gap in media and data literacy skills. It requires a lot of work to ensure that people of all ages and backgrounds have the skills and knowledge they need to navigate the complex world of media and data effectively. In this chapter, I explain the similarities and differences between these approaches and explore how they are currently being taught.

MEDIA LITERACY

Media literacy is accessing, analyzing, evaluating, and creating media messages in various forms, including print, audio, video, and digital content (*Media Literacy Defined,* n.d.). Media literacy is more important than ever in today's digital age, where information is readily available through various media channels. It's crucial to note that the media often portray different races, genders, and social classes in stereotypical or biased ways. Media literacy equips individuals with the tools to critically analyze these portrayals, questioning their origins and implications. Additionally, language can be a significant barrier in understanding and interpreting media content. Multilingual media literacy programs can help bridge this gap, making media literacy more accessible to people who speak different languages. Furthermore, it's essential for media literacy education to be culturally sensitive and inclusive. Educators should take into account the diverse backgrounds of learners to ensure that media literacy is not just a skill but a tool for social inclusion.

By being media literate, individuals can become more informed and responsible media consumers, able to examine and assess media messages and their sources critically. Media literacy can empower individuals in many ways. For instance, it can help them distinguish between fact and fiction, identify bias and propaganda, and recognize manipulative techniques used in media messages.

It can also enable them to understand better the cultural, social, and political contexts in which media messages are produced and consumed. Furthermore, media literacy can foster creativity and innovation by allowing individuals to express their ideas and perspectives through various media forms, such as writing, photography, video production, and digital media. In short, media literacy is a crucial skill enabling individuals to navigate the complex media landscape and make informed decisions about the media content they consume and create. By developing media literacy skills, individuals can become active and engaged media citizens, capable of participating in the media discourse and shaping the media culture.

CHALLENGES WITH MEDIA LITERACY

Some key challenges that need to be addressed include a lack of awareness. Many people need to be made aware of what media literacy is and why it is essential. There is a need for more education and outreach to help people understand the value of media literacy. There is a digital divide between those with access to technology and those without access (Taylor, 2022). This can limit the ability of some individuals to develop media literacy skills and can exacerbate existing inequalities. With the rise of fake news and misinformation, it has become increasingly challenging for people to distinguish fact from fiction. There is a need for more emphasis on critical thinking and fact-checking skills.

Several steps can be taken to promote media literacy. We need to ensure that media literacy is taught in schools and that students have the opportunity to develop these skills from a young age. Media literacy is not just the responsibility of educators but also of media professionals, policymakers, and parents. Collaboration between these groups can help ensure that everyone has access to accurate information and is equipped to navigate the media landscape. With the rapid pace of technological change, we need to be innovative in our approach to media literacy. This may involve the use of new technologies, such as virtual and augmented reality, to help people develop media literacy skills engagingly and interactively. In summary, media literacy is an essential skill for navigating the modern media landscape effectively.

While there are still many challenges to overcome, there are also many opportunities to promote media literacy through education, collaboration, and innovation. Media literacy and data literacy are connected in that they both involve critical thinking skills. In order to effectively analyze media content or interpret data, one must be able to ask questions, evaluate sources, and think critically about the information presented. Additionally, both skills require an understanding of how information is created, disseminated, and consumed in today's digital world.

DATA LITERACY

What is data literacy? There are many resources that explain data literacy, but one person who explains data literacy well is Tim Stobierski. Mr. Stobierski is a marketing specialist and contributing writer for Harvard Business School Online, who writes about data literacy. His compelling 2021 article explains: “Data literacy is a term used to describe an individual’s ability to read, understand, and utilize data in different ways. It doesn’t require an individual to be an expert—as a data scientist or analyst might be considered—but rather to show an understanding of basic concepts, such as different types of data, Common data sources, Types of analysis, Data Hygiene, Tools, Techniques, and Frameworks.” (p. 1).

Data Literacy is increasingly important today. It’s crucial to discuss the ethical implications of data collection, especially how it can disproportionately affect marginalized communities. Additionally, the potential for bias in data should not be overlooked. A lack of diversity in data science can perpetuate systemic inequalities, making it essential to address this issue in data literacy education. These issues are discussed in Chapter 1 as well as articles included in the appendix. Organizations and individuals are inundated with vast amounts of data. It empowers individuals to make informed decisions, identify trends, solve problems, and effectively communicate insights derived from data. Data can empower organizations and individuals to share information and collaborate. Data is to be shared and explored by individuals to gain skills and knowledge. ("The 2020 Global State of Enterprise Analytics", 2020)

With data being consumed by many people and platforms, individuals need to understand and analyze the data they encounter. Data literacy is not just about understanding numbers and statistics but also about being able to interpret and communicate the insights derived from the data. It involves understanding the context in which the data was collected, the biases that may be present, and how to use data to make informed decisions. In today's digital age, data is everywhere and being data literate is essential for success in many industries. Data literacy also involves understanding the ethical implications of collecting, analyzing, and using data and the importance of privacy and security. Overall, data literacy is an important skill for anyone who wants to be able to understand, analyze, and communicate insights derived from data.

SIMILARITIES

Media literacy and data literacy are related in many different ways. Media literacy looks at media and how it influences our reality. Data literacy is the understanding of how platforms like social media apps and beyond interact with society. It is essential to know that much of the information we encounter on a daily basis can be misleading. A better understanding of media and data literacy can help someone navigate through various media. Many sources out there are reliable, but when it comes to knowing which sources are reliable, it can become tricky. That is why it is essential to get a better understanding of media and data literacy. It's important to recognize the intersectionality of media and data literacy with issues of race, gender, and social class. Understanding how these literacies intersect with broader social issues can provide a more holistic approach to media and data literacy. Knowing even just a little about both can greatly improve one's knowledge of the media we all consume today.

UNDERSTANDING DATA LITERACY SKILLS

There are many benefits to developing data literacy skills, including the ability to identify patterns and trends, make data-driven decisions, and communicate insights effectively. Data literacy is becoming increasingly important in many fields, including business, healthcare, and education, as more and more organizations are relying on data to make informed decisions. For example, in healthcare, data literacy can empower community health workers in underserved areas to better understand and address the specific health needs of their communities. In education, teachers in diverse classrooms can use data literacy to tailor their teaching methods to better serve students from various cultural and linguistic backgrounds. With the growing importance of data, there is also a growing demand for individuals with strong data literacy skills.

Inclusive case studies featuring a diverse range of individuals and communities can further enrich the understanding of data literacy. These case studies can serve as practical examples that resonate with a broader audience. With strong data literacy skills, there are many career opportunities an individual can have. One example of such a career is a marketing analyst. Marketing analysts identify customer behavior, measure marketing campaign effectiveness, and optimize marketing strategies. Data literacy enables them to conduct accurate data analysis, segment audiences, track key metrics, and make data-driven recommendations for marketing decisions. Another job would be a data analyst. Data analysts play a crucial role in collecting, analyzing, and interpreting data to provide insights and support decision-making. With data literacy skills, they can effectively manipulate and analyze data, develop meaningful visualizations, and communicate data-driven findings to stakeholders.

Overall, data literacy is an essential skill for navigating the complex data landscape and making informed decisions. By developing data literacy skills, individuals can improve their understanding of data and use it to drive positive change in their organizations and communities.

In addition to the benefits of data literacy for individuals, there are also broader societal benefits. A more data-literate society can lead to better-informed decisions and policies, improved public health, and more effective and efficient use of resources. Data literacy can also address issues of inequality and social justice by providing insights into patterns and trends that may be affecting marginalized communities.

However, there are also challenges associated with data literacy, such as the potential for bias in data collection and analysis, the difficulty of interpreting complex data, and the risk of misinterpreting or misusing data. Individuals need to approach data with a critical mindset, be aware of the limitations of the data, and seek out diverse perspectives when analyzing and interpreting data.

Data literacy is not a static skill set but a constantly evolving one. As new technologies and data sources emerge, individuals must be willing to adapt and continue learning to remain data literate. Today data literacy is becoming an increasingly important skill for success in many fields and industries. It will continue to be essential for individuals and organizations to stay ahead of the curve.

MEDIA LITERACY AND DATA LITERACY SKILLS

Developing media and data literacy skills is crucial for navigating the complex media landscape and making sense of the vast amounts of available data. When learning about media and data literacy, people need to understand how they work together better. I argue that teaching media and data literacy should be taught so the student can learn how to navigate through them without concrete courses devoted completely to that topic. Instead, already existing courses can assign work that can help improve one's knowledge of media and data literacy. The student should be taught the overall understanding of media and data literacy so they can understand how to make decisions and identify misinformation.

To effectively teach media literacy and data literacy in schools in Massachusetts, teachers should have small assignments that help build media literacy and data literacy. Students with no knowledge of media literacy and data literacy would gain more information on the topics due to the small workload that is provided. In the process of these assignments, students and teachers can be more understanding and share their knowledge to further gain information. Massachusetts has also passed a bill that makes education for media literacy a high school graduation requirement. This also requires the Department of Elementary and Secondary Education to develop instructional guidelines in media literacy. The Department of Education has a working group to access and recommend revisions to policies and procedures on media literacy aligning with K-12 standards. This working group will consult with experts in media literacy including but not limited to academic experts and non-profit organizations. In the development of teaching and learning media literacy and data literacy, the Department of Education will assist in resources to aid and will provide and make sure media literacy and data literacy training opportunities are available.

Educators can incorporate these skills into various subjects, such as English, social studies, science, and math. Teachers can use real-world examples to demonstrate the importance of these skills and can provide opportunities for students to analyze media messages and data sets. Additionally, schools can offer classes or workshops specifically dedicated to teaching media and data literacy and can provide access to resources and tools that allow students to practice and develop these skills.

An example of how media literacy and data literacy could be taught is by engaging students in an interactive discussion about media and data topics. One example is encouraging students to share their perspectives and asking questions and then critically analyzing different media messages and data sets. Another example of how media literacy and data literacy could be taught is by doing a media analysis assignment. In doing a media analysis, students will evaluate and critically analyze different types of media. Further, the assignment can require students to identify the intended audience, the message it is sending, and any biases that may be presented. It is important to adapt the teaching methods to a specific audience, group, education level, and learning style of the students.

Incorporating media literacy and data literacy into the curriculum of schools in Massachusetts is essential for preparing students to be critical thinkers and responsible consumers and producers of information in our increasingly media-saturated and data-driven world and preparing students for their future career opportunities. Employers increasingly seek individuals with these skills to work in various industries, including media, marketing, and technology. Some other opportunities that focus more on data literacy are data science, data engineering, business intelligence analysis, and data journalism. These are just a few jobs that focus on their employees having data literacy skills. The demand for these roles continues to grow as organizations recognize the value of data in decision-making and innovation.

According to a 2021 report by Burning Glass Technologies (Bursin, 2021), a labor market analytics firm, media literacy skills were listed as a desired competency in job postings across a variety of fields, including journalism, public relations, advertising, marketing, and social media management. The report found that jobs requiring media literacy skills were growing at a rate of 6.5% annually. Equipping students from underserved communities with these skills through K-12 education can help level the playing field, potentially leading to a more diverse and equitable workforce. As these communities often face systemic barriers to employment in these growing fields, early media literacy education can be a step toward economic empowerment.

Additionally, having a solid foundation in media and data literacy can empower students to analyze and interpret information from multiple sources, leading to more informed decisions in their personal and professional lives. This is particularly impactful for marginalized communities, as being better-informed consumers and citizens can lead to more equitable access to opportunities and resources. Schools should also collaborate with experts in the field and consistently update their teaching methods to ensure students remain on the cutting edge of media and data literacy education. In turn, this comprehensive education will better position students to make valuable contributions to their chosen industries and communities, potentially leading to societal benefits such as a stronger, more diverse workforce and more equitable community development.

There can be many viewpoints on how we should teach media and data literacy, and many people might think there are better ways to teach literacy than what I have argued so far. Various opinions and criticisms exist regarding how these literacies should be taught. Below I consider a few of them and how they can be addressed.

Some argue that media literacy and data literacy are not essential skills and should not be a priority in education. They believe other subjects like math and science should take precedence. However, with the rise of fake news and misinformation, individuals need to be able to distinguish between credible and unreliable sources. When getting information from an unreliable source, it can cause the work to lose credibility. It also may cause insurrections due to the fact that misinformation or fake news is dangerous and can cause people to act violently (Hoʻoulu Staff, 2017).

Additionally, data literacy is essential for making informed decisions in various fields, including business, health, and politics. Critics argue that media literacy and data literacy are too complex for the average person to understand. They believe these literacies require specialized training and should be left to experts. While media literacy and data literacy can be complex, it is possible to teach them in a way that is accessible and understandable for the general public. Teachers can use real-world examples and hands-on activities to make these skills more tangible and relevant to students. While media and data literacy may require different approaches, they are interconnected skills. Understanding how to analyze and interpret data is critical in evaluating media sources and vice versa. Teaching these literacies in conjunction with each other is essential to provide students with a more holistic understanding of the information.

While there may be differing opinions and criticisms regarding how media and data literacy should be taught, it is essential to recognize the importance of these skills in today's information age. Teachers can use various strategies to make these skills accessible and relevant to students, including using real-world examples.

CONCLUSION

Media literacy and data literacy are essential and provide individuals with information and knowledge. Without media literacy, individuals would struggle to think critically about media and understand what's credible. Without data literacy, individuals would struggle to make data-driven decisions and communicate insights effectively. In today's world, these skills are extremely important and are vital in navigating the complexities of media and data effectively.

WRAP UP

Key Takeaways

- Media and data literacy are not just essential skills but also tools for social inclusion and empowerment, enabling individuals to make informed decisions and engage in societal discourse.
- The digital divide and systemic barriers like socio-economic status and educational opportunities can significantly impact access to media and data literacy, making education and outreach critical.
- Both media and data literacy are evolving fields that require ongoing education to keep pace with technological advancements;

innovative tools like virtual and augmented reality can enhance this educational process.

- Teaching media and data literacy is not just the responsibility of schools; it requires a multi-faceted approach involving educators, media professionals, policymakers, and parents to be truly effective.

Exercises

1. How do systemic barriers like socio-economic status, educational opportunities, and cultural factors impact access to media and data literacy in your community? Discuss specific examples.
2. Conduct a case study analysis of a media campaign or news story, identifying any biases, target audiences, and the techniques

used to convey the message. Discuss how media literacy skills could help someone critically evaluate this campaign or story.

3. Using publicly available data sets, perform a basic data analysis task, such as identifying trends or disparities in the data. You should also discuss any potential biases in the data and how data literacy skills can help them interpret the information.

REFERENCES

Bill S.213 188th (2013 – 2014). (n.d.). Massachusetts Legislature. Retrieved April 25, 2023, from https://malegislature.gov/Bills/188/S213

Bursin, J. (2021). *The Hybrid Job Economy*. https://www.burning-glass.com/wp-content/uploads/hybrid_jobs_2019_final.pdf

Confronting the Challenges of Participatory Culture: Media Education for the 21 Century. (2009). MacArthur Foundation. Retrieved April 4, 2023, from https://www.macfound.org/media/article_pdfs/jenkins_white_paper.pdf

Hobbs, R. (2010). *Digital and Media Literacy: A Plan of Action*. Aspen Institute. Retrieved April 4, 2023, from https://www.aspeninstitute.org/wp-content/uploads/2010/11/Digital_and_Media_Literacy.pdf

Ho'oulu Staff. (2017, January 7). The Real Consequences of Fake News – Ho'oulu. *The University of Hawaii Maui College*. https://maui.hawaii.edu/hooulu/2017/01/07/the-real-consequences-of-fake-news/

Jin, H. (2020, November 28). *How to Distinguish between Reliable and Unreliable Sources Online*. INKspire. Retrieved May 16, 2023, from https://inkspire.org/post/

how-to-distinguish-between-reliable-and-unreliable-sources-online/-MKNvPDSaxfUKjIneDIr

Media Literacy Defined. (n.d.). NAMLE. Retrieved April 4, 2023, from https://namle.net/publications/media-literacy-definitions/

Stobierski, T. (2021, February 23). *Data Literacy: An Introduction for Business*. HBS Online. Retrieved March 28, 2023, from https://online.hbs.edu/blog/post/data-literacy

Taylor, Kiara. "The Digital Divide: What It Is, and What's Being Done To Close It." Investopedia, April 15, 2022. https://www.investopedia.com/the-digital-divide-5116352.

"The 2020 Global State of Enterprise Analytics." Microstrategy, 2020. https://www.microstrategy.com/en/resources/research-and-reports/the-2020-global-state-of-enterprise-analytics.

PART V

THE AMERICAN MOTION PICTURE INDUSTRY AND BIG DATA

Figure 1. "A film reel being recorded by a film camera, digital art". 'Film Reel' by Brendan Smith, created with DALL-E-2

Chapter Written by Brendan Smith

Learning Objectives

- Explain how big data has transformed the film industry, particularly in predicting box-office success and targeting marketing campaigns.
- Gain an understanding of the ethical and equity considerations associated with using big data in the film industry, including the potential for data bias and underrepresentation of certain groups.
- Identify the advantages and limitations of using big data in various stages of film production and distribution, from pre-production to post-release analysis.

INTRODUCTION

The film industry has undergone a major transformation in the past decade, thanks to the emergence of big data. With the ability to predict box-office success for upcoming releases, film production companies can make more informed decisions about everything from advertising to release dates. This is a significant improvement from earlier methods such as focus groups and analyzing box-office receipts, and it's all thanks to the incorporation of social media and other new data sources. By using big data, companies can now analyze marketing campaigns and audience responses to find potential hits or even decide to delay or cancel a film. But that's not all; big data can also be used to predict award winners and to identify potential films that audiences will want to see. In this chapter, we'll explore the ways in which big data has revolutionized the film industry, and we'll take a look at its potential for further advancements in predicting box-office success. This is a behind-the-scenes look at how big data is used and even transforming the film industry. As well as the challenges and limitations of these approaches.

PREDICTING BOX-OFFICE SUCCESS IN THE FILM INDUSTRY

In the world of the film industry, there are many films that are released each year to the public and are successful, but there are also films that do not see the same success as the others. In the United States alone, 449 films were released in 2022. This number is slowly rising back to the 2019 level of 792 films after falling during the COVID-19 pandemic. (*U.S. & Canada*, 2023). Film production companies want to be able to figure out whether or not they should fund a project based on its predicted box-office success. Currently the methods that are being used are not officially known, but there is evidence that suggests that data is being used for predictions by the companies in the American Motion Picture Industry. For a film production company to be able to predict the success of a film when it is in the stages of pre-production, there needs to be something out there for the film, such as a trailer for the people to be able to see a preview of the project, before companies are able to begin making predictions. 20th Century Studios has machine learning models to help before pre-production, but there are no known results. With that being said, how and when can a production company within the American Motion Picture Industry predict the success of their films at the box-office? Read on to find out.

Before there was access to big data, companies used much simpler methods for prediction. Focus groups were employed to predict success for a film, as well as analyzing box office receipts to predict the potential success of a film based on similar films. There was one major issue with these methods, they all lacked easily accessible data, causing them to fail as reliable methods (Simon & Schroeder, 2019).

BIG DATA IN THE FILM INDUSTRY

Before continuing on, it is important to mention that the methods mentioned below are data sets and models created by third parties and not the companies within the American Motion Picture Industry. The third parties created models that they believe could be or could have been used by the companies. With the exception of the previously mentioned 20th Century Studios and Netflix, which will be talked about in detail later on, the companies in the industry tend to not release information about the models that they use. It can be assumed that this is the case due to the fact that the companies do not want competitors to be able to access or create similar models to them if they have seen relative success with their particular data sets and models. It is also important to note that, due to this being the case, we do not know from whom or where the data they use is collected from. This raises the question of potential data bias and how they are using such data.

The emergence of big data within the film industry has allowed for many avenues of box-office predictions to become available. But what is the big data within the film industry? The big data about the film industry used to be far more limited. It only held box-office receipts, surveys, focus groups centered around the awareness and the attitude towards a particular film, and the outlets in which advertising for films were placed (Simon & Schroeder, 2019). However, the amount of available data has increased significantly. One of the main contributors to this is social media. The data now includes the online views gained by various trailers for a film, the social media posts about a particular film and the positive and negative engagement that they garner in regard to the yet-to-be-released film (Gold et al., 2013, Simon & Schroeder, 2019). Another area in which data was gathered was from the comments of the website Rotten Tomatoes. Rotten Tomatoes is a website where users are able to view critical and audience ratings of films and television shows, as well as comments made by the audience in regard to the films. These different areas and aspects of the data now allow for much more robust ways for companies to be able to predict the success of their yet-to-be-released films and to plan accordingly for the potential success or failure.

How can the data be used by companies within the American Motion Picture Industry? With the emergence of new data, these companies can analyze the responses of the audience toward their marketing campaigns and to see if there is interest garnering in regard to the movie before its planned release. With these comes the most important prediction for the companies, the box-office success of their film (Simon & Schroeder, 2019). Being able to predict how well their movie will perform once it releases into theaters allows for the companies to have opportunities in front of them that may not have been present or usable without the knowledge of the data. These opportunities could include funding further advertising, delaying the release of the film, or even altogether canceling the film and its production.

HOW DATA HAS BEEN USED IN THE FILM INDUSTRY

Not all uses of the data go toward the success of the film at the box-office. The data can be used in a wide variety of ways. For example, in 2013, Farsite Group used data that they had gathered to predict the winners of six of the main awards for the 85th Oscars. They used Rotten Tomatoes ratings given by both critics and audience members, the box office success of each film, and if the films had won any awards at award shows that take place before the Oscars, such as the Directors Guild Awards, and the Golden Globes (Gold et al., 2013). In doing so, they were able to accurately predict 5 out of 6 of those awards. In 2014, Farsite Group once again predicted the Oscars (Pomerantz, 2014). There were no articles mentioning the outcomes of their predictions, but after cross referencing the Forbes article mentioning the predictions with the official Oscars website (*The 86th Academy Awards,* 2014), Farsite Group accurately predicted 6 out of 6 awards for that year.

Another example comes from a study done by Sitaram Asur and Bernardo A. Huberman (2013). Together, they analyzed 2.9 million tweets from 1.2 million different users about 24 different films. Considering the mentions of the film, the positivity of the tweets, etc. they were able to use a linear regression model that allowed for them to be able to show the relationship between the spread of people talking about the film and how successful the film was likely to be because of that. One example from their data set was the film *Avatar*. A week before its release it accumulated around 1212.8 tweets per hour. They were able to use this through their model to show that based on the number of tweets surrounding the film, it would be a successful film within its first week of release. Through their work, they were able to prove that data gathered from social media sites can effectively predict the future outcomes of a particular film's success at the box office. They were also able to prove that this method of analysis and prediction worked much better than the predictions of the Hollywood Stock Exchange. The Hollywood Stock Exchange is an online virtual stock market where users are able to buy and trade stock using virtual fake currency to make predictions for which movies will be a success at the box office. The method used by Asur and Huberman helped to prove that the more a film is positively talked about prior to its release, the better the film will perform at the box-office when it comes time for it to be released into theaters (Asur & Huberman, 2013).

An interesting tool that could eventually be used by many companies comes from 20th Century Studios. They have revealed that they began using machine learning models before they even started pre-production. These machine learning models collect data and help find potential films that audiences will want to see. 20th Century Studios uses this to guide themselves when buying a script (Kapoor, 2021). They take labels created for their films and then they feed those through the machine learning models to help them discover potential scripts. This changes how the process works. Traditionally, a producer would have assistants who go through the scripts for them and author short reports on the scripts. It is then up to the producer to read the reports and decide which film they would like to make next. The machine learning models can now choose the scripts and then the assistants narrow down the chosen scripts rather than narrowing down every script. This could increase the output of more successful films overall. It can be seen as a potential huge money saver for companies looking to produce certain films. A lackluster script can be better avoided rather than be made and create a net loss after it has been produced into a movie, allowing for there to be less of a net loss when that film has a high budget and ends up not performing well at the box-office and after word spreads that there is not much substance to the film due to the script. The model considers what audiences might want to see next, but it is unable to account for any unexpected breakthroughs in the industry. This is due to the model relying on the earlier scripts that are considered successful by the studio.

"Predicting Movie Prices Through Dynamic Social Network Analysis" used both the Internet Movie Database

(IMDb) and the Rotten Tomatoes moving rating parameters, along with the "buzz" around a film and posts gathered from the IMDb forums (Simon & Schroeder, 2019). This allowed them to be able to make predictions for a film during the first four weeks after its release. But while they had some success with this method, the information does not really become useful for the studios as their film is already in theaters and they can figure how well their film will do based on their opening weekend.

Another study that garnered results came from, "Predicting consumer behavior with Web search" (Goel et al., 2010). They were able to use data gathered from Yahoo!'s search engine for box-office predictions (Simon & Schroeder, 2019). The data that they gained from the search engine came in the form of searches from individual users. They were able to compile the data based on whether or not there was a link to IMDb within the immediate search results and they then mapped out the movies based on which movie the IMDb link led to. With this, they were able to create predictions that worked well. They also found that the results worked particularly well due to users using the search engine to search for the film they were interested in and where they would eventually be able to see the film near them when it was released into theaters. But even still they pointed out an issue with this method, "the main advantage of using a search behavior may not be accurate but rather the ready availability of these data." (Simon & Schroeder, 2019, p.554). Having the data sets there to use is indeed a fantastic advantage due to their availability, but having Simon and Schroeder say that they may not be the most accurate leads to the conclusion that those data sets would be most beneficial if they were a part of a model that takes data from multiple sources to be able to accurately create box-office predictions by analyzing multiple data sets from different sources rather than relying on the data from one particular source. This raises important equity considerations. For instance, the data could be skewed due to the digital divide, underrepresenting people from lower socio-economic backgrounds who may not have regular internet access. Additionally, the data may carry cultural and language biases, as it predominantly captures the behaviors and preferences of majority populations or those who are more active online.

INTO THE FUTURE

New ways of using data are regularly being invented and used. Surveys and categorization of age ranges that are used in data are becoming more finely sharpened. “That blunt instrument is fast giving way to computers that can render us in fine detail by picking up the trails of digital breadcrumbs we leave online and building them into predictive models of what we like and don’t like.” (*Big Data and Hollywood*, n.d.). The methods allow for a much more exact prediction model as to who will want to see which movie and how to get that movie to appeal to an even wider audience. “Studios can use these real-time opinion assessments to do all kinds of tweaking after a movie has been made; targeting specific demographics in marketing campaigns, tailoring trailers so that they appeal to the kinds of people who will be drawn to a particular movie, pushing distribution to geographic areas where the target audience lives.” (*Big Data and Hollywood*, n.d.). implementing these into the film industry will allow a film company to take a hold of their box-office success to an extent. This will allow those companies to obtain a much wider audience for their films. And in doing so, they will be able to increase their box-office success in ways that were previously not available to them.

Another potentially game changing system comes from

Cinelytic. They have been working on an AI system that aids film production companies in new ways. “It licenses historical data about movie performances over the years, then cross-references it with information about films’ themes and key talent, using machine learning to tease out hidden patterns in the data.” (Vincent, 2019). This essentially creates an AI producer for companies. Analyzing all aspects of the data allows for it to look for those trends that work and when they may work and when to shift the focus to try to gain attention from a wider audience.

A problem also arises with the use of AI in film. For example, say that it was to be used to gather audience feedback before the film was released. It could then make recommendations that can be used to make the film more in line with what the audience is expecting it to be. For example, before the film *Snakes on a Plane* was released, it began to garner attention and so the studio decided to have reshoots for parts of the film so that they could incorporate feedback from the audience (Simon & Schroeder, 2019). And upon doing so, when it came time for the theatrical release of *Snakes on a Plane*, it fell short of what it was predicted to make at the box office. Taking this into consideration, while this is speculatory, if the AI is to take into account what this audience is interested in seeing based on their feedback, what is there to stop the method from being the same in terms of failure when relating this method to the one used for *Snakes on a Plane*?

THE ADVANTAGES OF BIG DATA IN THE FILM INDUSTRY

One of the advantages of big data is that it allows for a "...more accurate and detailed customer information at the individual level and uses the information for a very narrow and specific segmentation of customers..." (Rust & Huang, 2014, p.209). This was mentioned earlier with 20th Century Fox. It also allows for a better and more exact picture of the audience so that they become more understood by those who are creating films. This can allow for there to be a much more diverse and better representation of various groups of individuals who were once unrepresented or misrepresented within film.

> *"Various sources of data can be combined, including not just social media data but also geo-location (where which movies are popular), credit card data, and the like." (Simon & Schroeder, 2019, p.558).*

Having geo-location available is an essential part of the process for film production companies. Being able to know where certain genres are more popular than others allow for the companies to better plan out and spend money on the number and locations of movie theaters where the film is released. Being able to limit how many movie theaters play a film in a certain location allows for money to be saved rather than be spent on too many theaters in areas where that film genre may not be as successful as it is in other regions. However, this approach raises ethical concerns, particularly for rural or economically disadvantaged areas that may not have the customer base to warrant the release of certain films. These areas could be left out, limiting their access to diverse cultural content. Moreover, the data could perpetuate existing biases, as big chains might focus only on genres that are already popular in specific geographic locations, thereby reinforcing existing cultural divides.

When a studio has their films shown at a chain like AMC, it tends to play in most if not all of the locations that that chain owns. But for smaller independent theaters, they don't always have the ability to showcase the films in their cinemas. Data could help with this to an extent. In a location where a major cinema chain isn't present, and a smaller cinema is, then if that area matches the demographic of those who have interest in a film, a studio can make a deal with that smaller independent cinema to have their film shown at that cinema to increase profit in the area. Allowing for a film to only be shown in theaters in locations in which those genres of films are popular amongst movie goers allows for there to be a better box office return for the companies, rather than if they were to showcase their film in theaters where the genre of their film is less popular than other genres.

THE LIMITATIONS OF BIG DATA IN THE FILM INDUSTRY

One of the limitations is the lack of data on low-budget independent films (Simon & Schroeder, 2019). It becomes harder to successfully predict the box-office success of films that are smaller in scale. This is due to the lack of information in regard to the film. The less information that is available in a film, the harder it becomes for a prediction model to be able to use the data to accurately make a prediction.

Another limitation is that it does not consider who the potential audience for the film is and the audience for its actors. For example, the film *Ticket to Paradise* was not expected to do well in the United States (U.S.). Upon its release in the U.S. It had made $80 million overseas and was predicted to only accumulate $6.4 million in its opening week in the U. S. But in a surprise, *Ticket to Paradise* managed to secure $16 million in its U.S. debut (McClintock, 2022). It was reported that sixty-four percent of the audience for the film in its first weekend were older than 35 (*What the "Ticket to Paradise" Box Office Opening Says about the State of the Rom-Com*, 2022). The age of the audience is a key factor in this. In a movie starring two older actors who are less known to the younger audiences, the main audience who is likely to go see the actors and actresses are familiar with those actors and actresses from when the audience themselves were younger and watch their films.

There are no officially known prediction models, except for one exception, which are able to make predictions for which actors should be hired for a particular project. The one exception to this is Netflix. In the past it has been revealed that for the Netflix series *House of Cards*, Netflix was able to use data gathered from the streaming platform to make a prediction that pairing David Fincher and Kevin Spacey would be a success. This was based on the popularity of films directed by David Fincher and films starring Kevin Spacey (Carr, 2013).

When it comes to predicting success at the box-office, there is one limitation that can't be accounted for. With predictions based upon social media users' activity in regard to the film before its release, there is no way to accurately predict the potential flop of the film at the box office until it has already happened. A film can be predicted to be a success with all the talk surrounding it prior to its release and have the estimated box-office receipt and how many theaters it will play in, but nothing can account for the word of mouth spread of negativity towards a film that is not rated well in the eyes of the audience. The audience who has seen it and disliked it overall can lead to the unpredictable possibility of the film becoming a failure at the box-office when it was previously seen as a potential enormous success. Asur and Huberman (2013) have found that after the release of a film, sentiment on social media can affect the predictions of a film's box-office revenue.

Another limitation for big data within the industry comes down to when the usable models can create accurate predictions for the films. "Big Data Goes to Hollywood: The Emergence of Big Data as a Tool in the American Film Industry" brings this to light. Asur and Huberman's (2013) Twitter-based model was only able to make a prediction that could be seen as dependable the night before the film was released into theaters in the United States (Simon & Schroeder, 2019). While the model has shown that it can help to accurately predict which films will be a success, there is no real help when it can only supply the info the night before. At that point, it does not become beneficial for the studio to know at that point due to there being nothing for the studio to do about how the film is distributed.

A main point to remember from the Yahoo! and other search engines includes: when solely working off of the data sets from only one or a few sources at a time can lead to success with predictions, those predictions will contain a lot more failed predictions than there would be if you were to utilize many various data sets to create an overarching model in which box-offices predictions can become more and more successful. One major limitation of this analysis is that companies reveal little to no information on their prediction models, but we do know that they create these overarching models to consider the possibilities from all sides (Kapoor, 2021).

OUTLOOK

In "Big Data Goes to Hollywood: The Emergence of Big Data as a Tool in the American Film Industry", Felix M. Simon and Ralph Schroder (2019, p. 560) bring up a valuable point: "...rationalization has of course affected the kinds of movies that are made and how they are made. And films with large budgets are affected more than lesser movies simply because the stakes are so high." Essentially, what is being argued is that while those working for studios claim that there is no effect upon the creative process of a film, there still is much effect that it has. One of the reasons a person writes a script or wants to direct a film is to use their creativity. But there comes a point at which that creativity begins to become hindered. And big data can be seen as a contributing factor to that hindrance. This can especially be seen with 20th Century Studios' models and Cinelytic's AI model. If a script is not seen as appropriate for what should be made, then no chance is taken on that script, or that script is then changed to meet the algorithmic recommendations rather than allowing for the full creativity of its creators. And, for the AI, if it looks for hidden patterns that will help make a film more successful, the studio would benefit and the creativity of the film would be hindered. If the film is to be changed to be more in line with what is supposedly the best way to make it a complete success, then it can fail in the creative process as it tries to stick to a formula that is predicted to work. The use of big data in the film industry is beneficial when it comes to marketing in advertising as mentioned throughout this chapter, it is only when it enters the pre-production world that it becomes a moral question of creativity.

CONCLUSION

The film industry has evolved in the past decade with the emergence of big data, allowing film production companies to predict box-office success for their upcoming releases. However, it's crucial to question whether these data-driven methods are inclusive and equitable. For instance, the #OscarsSoWhite campaign highlighted the disparities in awards and recognition across different populations. The predictive models being used may replicate existing biases. Therefore, while this is a significant improvement from the earlier methods such as focus groups and analyzing box office receipts, problems remain The incorporation of social media and other new data sources has revolutionized the industry, allowing for more in-depth analysis and prediction of the success of upcoming films. Companies can use this data to analyze their marketing campaigns and audience response, leading to more informed decisions about advertising, delaying, or even canceling a film. However, as the industry leans more into data-driven decision-making, it's essential to ensure that these methods don't perpetuate existing inequalities or overlook diverse talents and stories. Additionally, big data can be used to its full extent through the numerous ways in which it can be used throughout the entirety of the American Film Industry. The film industry has taken a big step forward, thanks to big data and its potential for further advancements in predicting box-office success, but it must also take steps to ensure that this progress is inclusive and equitable.

WRAP UP

Key Takeaways

- Big data has revolutionized the film industry by enabling more accurate predictions of box-office success, thereby informing decisions on everything from marketing to release strategies.
- While big data provides valuable insights, it also raises ethical concerns such as data bias, which can perpetuate existing inequalities and overlook diverse talents and stories in the film industry.
- Traditional methods like focus groups have been largely supplanted by big data analytics, but limitations still exist, such as the scarcity

of data on low-budget independent films and the late timing of reliable predictions.

- The application of big data is extending into various aspects of film production, even influencing script selection and casting decisions, which raises questions about the balance between data-driven decisions and creative integrity.

Exercises

1. How does the use of big data in the film industry compare to its use in other industries like healthcare, retail, or finance? Discuss both the advantages and ethical concerns that are unique to the film industry.
2. Consider a recent film that either succeeded or failed at the box office. How might big data

analytics have influenced the film's marketing strategy, release timing, or even content? Provide specific examples to support your analysis.

3. Debate the implications of using big data to influence creative decisions in filmmaking, such as script selection or casting. Do you think the use of data analytics enhances or hinders artistic creativity? Justify your position.

REFERENCES

2014 | Oscars.org | Academy of Motion Picture Arts and Sciences. (n.d.). Retrieved April 18, 2023. https://www.oscars.org/oscars/ceremonies/2014

Asur, S., & Huberman, B. A. (2013). Predicting the Future with Social Media. *Applied Energy*, *112*, 1536–1543. https://doi.org/10.1016/j.apenergy.2013.03.027

Big Data and Hollywood: A Love Story. (n.d.). Theatlantic.Com. Retrieved April 4, 2023. https://www.theatlantic.com/sponsored/ibm-transformation-of-business/big-data-and-hollywood-a-love-story/277/

Carr, D. (2013, February 25). Giving Viewers What They Want. *The New York Times*. https://www.nytimes.com/2013/02/25/business/media/for-house-of-cards-using-big-data-to-guarantee-its-popularity.html

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(41), 17486–17490. https://doi.org/10.1073/pnas.1005962107

Gold, M., McClarren, R., & Gaughan, C. (2013). The Lessons Oscar Taught Us: Data Science and Media &

Entertainment. *Big Data*, *1*(2), 105–109. https://doi.org/10.1089/big.2013.0009

Kapoor, A. (2021, November 3). *The Role of Data Science In Movie Making*. The Smart Cube. https://www.thesmartcube.com/resources/blog/datawatch-lights-camera-analytics-the-secret-role-data-plays-in-movie-making/

McClintock, P. (2022, October 22). Box Office: 'Black Adam' Heads for Solid $60M Opening While 'Ticket to Paradise' Travels in Style. *The Hollywood Reporter*. https://www.hollywoodreporter.com/movies/movie-news/black-adam-box-office-previews-ticket-to-paradise-1235246172/

Pomerantz, D. (n.d.). *Can Social Networking Accurately Predict Oscar Winners?* Forbes. Retrieved April 18, 2023, from https://www.forbes.com/sites/dorothypomerantz/2014/02/24/can-social-networking-accurately-predict-oscar-winners/

Rust, R. T., & Huang, M.-H. (2014). The Service Revolution and the Transformation of Marketing Science. *Marketing Science*, *33*(2), 206–221.

Simon, & Schroeder, R. (2019). Big Data Goes to Hollywood: The Emergence of Big Data as a Tool in the American Film Industry. In *Second International Handbook of Internet Research* (pp. 549–567). Springer Netherlands. https://doi.org/10.1007/978-94-024-1555- 1_63

U.S. & Canada: Movie releases per year 2022. (2023, February

13). Statista. https://www.statista.com/statistics/187122/movie-releases-in-north-america-since-2001/

Vincent, J. (2019, May 28). *Hollywood is quietly using AI to help decide which movies to make*. The Verge. https://www.theverge.com/2019/5/28/18637135/hollywood-ai-film-decision-script-analysis-data-machine-learning

What the "Ticket to Paradise" box office opening says about the state of the rom-com. (2022, October 23). Los Angeles Times. https://www.latimes.com/entertainment-arts/movies/story/2022-10-23/ticket-to-paradise-box-office-julia-roberts-george-clooney

PART VI
DATA IN SPORTS MARKETING

Chapter Written by Sophia Moore[1]

Learning Objectives

- Understand the various sources and types of data used in sports marketing, including community data, private organizational data, and public data repositories.
- Gain insights into the ethical considerations surrounding the collection and use of big data

1. This chapter was written with assistance from ChatGPT.

in sports marketing, particularly in relation to audience privacy and personalized marketing.

- Analyze and discuss the future implications of Artificial Intelligence (AI) in the field of sports marketing, including its potential for predictive analytics, real-time insights, and ethical concerns.

INTRODUCTION

Today, marketing has become part of nearly every business. Through marketing, businesses and companies are able to engage with their consumers, which allows them to increase sales and their brand's identity (Emeritus, 2022). A major part of marketing is reaching your target audience. In order for companies to understand who exactly their target audience is, big data begins to play a huge role. Data such as advertisement engagement, sales, audience demographics, and more help businesses understand how to market their brand better.

With all this data being collected on individuals to better understand what meets their needs and wants, there is an ethical line that must be drawn. Of course, data is essential and almost unavoidable for an effective marketing campaign, but this data should not breach a user's privacy rights. Later in this chapter, the topic of ethically right and wrong uses of data used for marketing will be analyzed.

In this chapter, the importance of big data used specifically for sports marketing will be discussed. Sports marketing is an industry that relies heavily on data analysis to help understand consumer behavior and drive better business decisions. Data overall has become a key component in sports marketing, as it helps teams, leagues, and sponsors identify key trends, understand audience demographics, and make informed decisions about marketing strategy.

UNDERSTANDING BIG DATA

The technical definition of big data is “high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation,” (“The role of Big Data in sports marketing,” 2023, para. 3). Big data benefits sports marketing by providing a wide set of data; with a wide range of data, more effective business strategies can be built. These business strategies not only can benefit the business but also the customers, employees, and players.

Effective sports marketing with the assistance of data can help a player gain more exposure and reach a wider audience. This can lead to increased recognition and a higher profile, which can translate into greater endorsement deals, better contracts, and greater opportunities for career advancement. Fans can also experience benefits such as better ticket pricing and content that caters to their specific needs.

Big data allows for a comprehensive understanding of the audience, even in industries with diverse and vast audiences, such as the sports industry. This understanding encompasses their traits, attitudes, and actions, enabling the prediction of future actions. In essence, big data provides a precise and continually evolving map of the target audience, which can be used to tailor actions, timing, and targeting strategies to effectively reach them ("The role of Big Data in sports marketing," 2020).

SOURCES OF DATA IN SPORTS MARKETING

Sports stand to gain significant benefits from the increased availability of community data through blogging sites, social network trends, and content communities. This is evidenced by the fact that, in 2019, three of the top ten most searched stories on the Internet were related to sports (Moreno, 2019). In response, sports leagues are increasingly turning to digital programming to engage and connect with fans. Digital programming is essentially media found online such as websites, social media, and sports apps. The majority of sports fans interact on social media, which demonstrates the platform's effectiveness in facilitating engagement. Iconic soccer teams such as Real Madrid and Barcelona have amassed over 210 million Facebook and Twitter followers each, emphasizing the reach and impact of social media platforms in the world of sports. Brock and Khan (2017) have identified social media as a vast source of big data that can be utilized to derive insights and inform decision-making in sports.

The second development in sports data involves the utilization of private data collected within sports organizations through recording consumer transactions. This is achieved through the use of clickstream to categorize visits by location, purchase, search, and the use of mobile applications. Clickstream data is the information collected about a user while they browse through a website or use a web browser (Gillis, 2022). The emergence and popularity of technology have contributed to the digitalization of sports, granting access to customers' and firms' digital interactions (Lazer and Radford, 2017). This interconnectedness of sports leagues and customers through digital platforms allows teams and sponsors to gather and share unique information. DeSchriver et al. (2021) provide an example of the potential benefits of this approach by collecting daily performance data on hotel occupancy rates and average daily rates to examine the impact of Southeastern Conference college football games on local hotel demand from 2003 to 2017. This study collected data from the hotel management analytics firm Smith Travel Research (STR), which represents proprietary data in the tourism and hospitality industry. The study highlights the use of volume and variety data to advance demand literature in sport and hospitality.

The third area of data availability in sports comes from a variety of sources such as government, research funding agencies, professional societies, universities, individual researchers, and other public data repositories. Uhlir and Schröeder (2007) have suggested that publicly funded data can be beneficial for reuse by a broad range of researchers, socioeconomic applications, and the general public. Large datasets related to sports have been published by the National

Collegiate Athletic Association, sports-reference.com, and official university websites. For instance, Jensen et al. (2020) collected 169,479 observations to explore donor behavior in the intercollegiate athletic industry. The study found that the probability of donor contraction increases with decreasing economic growth. Additionally, public data repositories offer historical and other data for professional and Olympic sports, as well as all co-educational postsecondary institutions that receive Title IV funding and have an intercollegiate athletics program. Overall, sport marketing researchers can benefit significantly from using public data produced by various entities to enhance the quality and productivity of research. (Mamo et al., 2021).

While there has been progress in collecting data from community, private, and public sources, there are still many opportunities for continued contributions. Sports blog websites and fan-hosted podcasts are emerging as potential sources of valuable data that can help sport marketers uncover fans' opinions and create new business opportunities. This data on fans' opinions can be gathered through comments and deciphering the positive from the negative feedback through tools such as sentiment analysis. The data collected can also be expanded from text data to image, audio, and video data from various sources (Mamo et al., 2021). This would allow researchers to gain a deeper understanding of the role that different forms of media play in the world of sports marketing.

DATA AND TARGET AUDIENCE

Sporting events bring in a massive amount of viewers. For example, the 2023 Super Bowl brought in an estimated total of 113 million viewers, while an estimated 1.5 billion people viewed the 2022 World Cup (Nielson, 2023). With such a huge audience, it may seem overwhelming to think about how franchises are supposed to cater to their needs. This is where data on audience demographics becomes essential.

To break down data demographics, let's use the world's most popular sport as an example: soccer. In a study found on Doxee ("The role of Big Data in sports marketing," 2020), the Italian soccer team, Juventus, found a change in their audience. Until a few years ago, Juventus had a fan base that consisted mainly of males, mostly from Italy or Europe. However, through data polled from social media following and engagement, it was found that more and more women have begun to follow soccer. More and more people who live outside of Europe and Italy also follow Juventus. Juventus' star player, Cristiano Ronaldo, who has 241 million followers on Instagram compared to Juventus' following of 38 million, has a demographic of 61.5% male and 38.5% female (Starngage, n.d.). This trend of more female engagement in sports has become common throughout athletics. To find the data of this demographic, social media has become a big tool. Through social media, data such as followers, liked posts, and views can be broken down into demographics.

So what does this data mean for future sports marketing? Using data to understand a demographic allows sports marketers to make their franchise more personal. For example, from data showing a trend in women watching sports, franchises can respond by selling more merchandise made for women. This could mean merchandise made in women's sizes or more popular items for women such as leggings.

Another way data is collected for the target audience is through ticket sales. As Pete Giorgio (n.d.) has written: "With richer data, sports teams can know who was at the game, their in-stadium purchase history, and where they moved within the stadium. Having this specific information will enable more focused sponsor targeting and authentic engagement both inside and outside the stadium" Combing these metrics will allow a much deeper analysis of how fans are interacting with products and advertisements.

DATA USED FOR REVENUE

Not only do sporting events get a massive engagement through media, but in-person sales also generate a large amount of revenue. The average crowd size for an NFL game in 2022 was 69,442 people (Smith, 2023). To guarantee sports fans will be more likely to attend games, it is vital to set prices at an ideal price. Pricing is very important because it needs to bring in profit and be realistic for fans, the ideal price of a ticket can be found through data.

Researchers have created a tool to optimize sports ticket prices for both management and fans, in order to gain a better understanding of what people are willing to pay for sports tickets. This innovative approach is designed to improve the management of ticket prices and make them more attractive to fans, while also ensuring that they are fair and reasonable. This research was explored through Manufacturing & Service Operations Management by co-authors Robert Easley, John W. Berry Sr. Department Chair and Professor of Information Technology, Analytics and Operations at the University of Notre Dame, and Ovunc Yilmaz, assistant professor at the University of Colorado Boulder (Wampler, 2022).

The research team collaborated with an NCAA Division

I football program and analyzed its ticket sales data (Greene et al., 2017). Through a careful review of fans' purchasing behaviors and demographics, the researchers investigated two primary sales channels: season tickets and single-game tickets. By scrutinizing the audience segments within each channel, the research team was able to identify and differentiate between different groups of fans based on their purchasing patterns and preferences. This approach allowed for a more nuanced understanding of how fans engage with the ticketing process and how the program can better tailor its ticketing strategies to meet their needs.

Through this research, season ticket holders were broken down into three categories: big donors, the public, and employees. General ticket sales were also broken down into three categories: donors, alumni, and parents.

An unexpected discovery from the analysis of customer segment data was that as the number of available seats in a given section dwindled below a certain threshold, fans appeared to be less interested in purchasing seats in that section. This trend may suggest that fans do not perceive the remaining seats, which are typically located on the fringes of the section, to be a good value for their cost. Moreover, the research uncovered a divide among fans in terms of their price sensitivity: some fans prioritized watching the game from the best seats, regardless of the cost, while others were content with watching the game from the cheapest seats available. These insights shed light on the complex factors that influence fans' ticket purchasing behaviors and can inform more targeted and effective ticket pricing strategies (Wampler, 2022).

DATA USED FOR CAMPAIGNS

Advertising campaigns are fundamental for marketing. They allow companies to capture the attention of their core audience. Campaigns can be done in multiple ways and reach their audience through multiple platforms. The key for a successful campaign is knowing which platform will reach their audience the best, this is where data is important. Data can identify which platform the target audience engages with most and use that information to create the most effective campaign.

There are a few brands that are immediately associated with sports when they are mentioned. These are brands such as Nike, Addias, and Under Armour that have been notorious for their collaborations in the sports industry. One of the most common sports marketing strategies is using popular players to promote products. Data is essential for this marketing approach because brands need an understanding of what individual will be well perceived by their target audience.

In 2019, Nike collaborated with former NFL player Colin Kaepernick to promote athletes following their dreams. This was a campaign that was very strategic and took an immense amount of data to predict audience reactions considering a controversy surrounding the player at the time, which we will begin to explore further. Kaepernick began kneeling during the pre-game anthem in 2016 as a way to protest racial injustice (The Guardian, 2019). Despite negative responses from this campaign, including distasteful comments from former president, Donald Trump, the campaign proved to be successful with a resulting 5% increase in Nike stock.

In this particular campaign a variety of data has to be collected considering all perspectives. Data also has to be collected on how advertising can affect culture and individuals. This data is shown through audiences' positive and negative comments on social media posts and their engagement with the brand, such as the stock increase in Nike.

A study was done via questionnaires by Yun Kuan of Portland University in 2019 to get a better understanding of how audiences perceive controversial campaigns. In order to understand how controversial advertising impacts consumers' views on the social issue and the brand, participants were asked to react to two advertisements by different brands featuring different controversial issues. This was done to avoid bias on the specific issue or brand. "A major factor of this study includes participant's political views and past life experiences; therefore, an internet-mediated questionnaire was utilized to obtain participants from throughout the United States, thus ensuring a variety of backgrounds," (Yun Kuan, 2018, p. 20).

Through this research, the data revealed that controversial advertising is similar to a mirror, as it reflects individuals' pre-existing values. The International Advertising Association (1977) explains that controversial advertising is attractive to individuals who already support the cause and thus, it mirrors their political and social perspectives. If controversial advertising is a mirror, then corporations may not be effective in using it to alter people's political and social opinions (Yun Kuan p. 32).

This shows the amount of thought Nike had to put into creating campaigns. Data must be collected to represent multiple different perspectives before producing any new marketing strategies. This is because sports bring in a very diverse audience that have many different views.

THE FUTURE OF SPORTS MARKETING

Artificial Intelligence (AI) is already changing the sports marketing landscape in numerous ways, and is poised to have an even greater impact in the future. The impacts include more personalized marketing, predictive analytics, real-time insights, and influential marketing.

AI technology enables marketers to collect and analyze vast amounts of data, which can be used to personalize marketing messages and promotions to individual fans. This data consists of the topics mentioned prior in this chapter, but, with AI, it will be gathered more efficiently. This will include helping identify the most effective social media influencers to partner with based on factors such as audience reach, engagement, and demographics. This can create a more engaging and relevant experience, leading to increased fan loyalty and engagement.

AI-powered predictive analytics can be used to forecast future trends in fan behavior, identify opportunities for growth, and optimize marketing campaigns for maximum impact. AI algorithms can also provide real-time insights into fan engagement, sentiment, and behavior, allowing marketers to make rapid adjustments to campaigns and messaging.

Since AI is able to collect a vast amount of data at a fast rate, fans could experience a lot of personalized marketing in the future. An example of this is sports apps, such as the ESPN mobile app. When a consumer creates an account, AI can be used to collect data on this person. This could result in articles relating to the consumers favorite teams or players being promoted more in their feed. It could also result in advertisements being promoted for tickets to their favorite teams games or promotions for their favorite players merchandise.

For some consumers, the idea of more personalized marketing may seem appealing because their interests will be met. However, some may find an ethical concern with the amount of data that is being collected about themselves. There is a fine line between data being collected for the improvement of marketing and invading users privacy. The future of AI is still mysterious since it is still an evolving technology, but it is essential that its use is ethical and fair.

CONCLUSION

Sports marketing is of paramount importance due to its ability to generate significant revenue streams. Sports have a vast and passionate fan base, and effectively marketing sports events, teams, and merchandise can lead to increased ticket sales, merchandise purchases, and sponsorship deals. These revenue sources support the sports industry and contribute to local economies, creating jobs, and driving tourism. Furthermore, sports marketing plays a crucial role in building and maintaining the brand image of teams, athletes, and sporting events. Successful marketing campaigns can enhance brand visibility, recognition, and reputation, leading to increased fan loyalty and attracting new audiences.

Data has become a valuable asset in sports marketing, providing several benefits that contribute to its effectiveness and success. Data allows sports marketers to gain deep insights into their target audience. By analyzing demographic information, consumer behavior, and preferences, marketers can understand their fans' interests, motivations, and consumption patterns. This knowledge helps in creating personalized and targeted marketing campaigns that resonate with specific segments, leading to higher engagement and conversion rates.

Data also enables sports marketers to measure and track the effectiveness of their marketing efforts. With the availability of analytics tools and platforms, marketers can collect and analyze data on key performance indicators such as advertising campaign effectiveness, social media engagement, ticket sales, and merchandise purchases. This data-driven approach allows marketers to identify what strategies and channels are delivering the best results, enabling them to optimize their marketing campaigns and allocate resources more effectively.

The future of sports marketing will continue to improve and become more effective with the assistance of new AI technologies. Data will be able to be collected at faster rates resulting in more marketing strategies. Of course, as more data is collected more quickly, it will be difficult to ensure it is unbiased. There will also be an increase in personalization in marketing that will fit the interests of individual fans. The industry is constantly evolving along with technology and continues to change and shape the sports world.

WRAP UP

Key Takeaways

- Big Data plays a critical role in sports marketing by providing deep insights into audience demographics, preferences, and behaviors, enabling more targeted and effective marketing campaigns.
- Ethical considerations are paramount when collecting and using data for sports marketing to ensure that user privacy is respected and that the data is used responsibly.
- Advances in technology, particularly Artificial Intelligence (AI), are poised to revolutionize sports marketing by enabling real-time insights, predictive analytics, and highly personalized marketing strategies.

- Various sources of data, including community blogs, private organizational records, and public repositories, provide a rich set of information that can be leveraged to make informed decisions in sports marketing.

Exercises

1. How does the ethical use of data in sports marketing align or conflict with your personal views on privacy? Discuss the balance between effective marketing and consumer privacy.
2. Conduct a mini case study on a recent sports marketing campaign that utilized big data or AI. Analyze the campaign's effectiveness, ethical considerations, and the types of data used.

3. How do you envision the role of Artificial Intelligence in the future of sports marketing? Consider both the benefits and potential drawbacks, such as issues of data privacy or bias.

REFERENCES

Brock, Vitor, and Habib Ullah Khan. "Big Data Analytics: Does Organizational Factor Matters Impact Technology Acceptance?" *Journal of Big Data* 4, no. 1 (July 10, 2017): 21. https://doi.org/10.1186/s40537-017-0081-8.

DeSchriver, Timothy D., Timothy Webb, Scott Tainsky, and Adrian Simion. "Sporting Events and the Derived Demand for Hotels: Evidence from Southeastern Conference Football Games." *Journal of Sport Management* 35, no. 3 (May 2021): 228–38. https://doi.org/10.1123/JSM.2020-0268.

Emeritus. (2022, April 22). What is the importance of marketing for businesses? discover the undiscovered. Emeritus. https://emeritus.org/blog/what-is-the-importance-of-marketing-for-business/#:~:text=The%20importance%20of%20marketing%20for%20your%20business%20is%20that%20it,%2C%20reputation%2C%20competition%2C%20etc.

Giorgio, Pete. "How Data Can Help Drive Sports Sponsorship and Fan Engagement." Deloitte United States, n.d. https://www2.deloitte.com/us/en/pages/consumer-business/articles/fan-engagement-analytics-improve-fan-experiences.html.

Gillis, A. S. (2022, May 5). What is clickstream data (Clickstream Analytics)?. Customer Experience. https://www.techtarget.com/searchcustomerexperience/definition/clickstream-analysis-clickstream-analytics

Greene, Amanda, Kason O'Neil, and Gary Lhotksy. "Exploring a New Division 1 Football Program on a University Campus: An Application of Collaborative Action Research in Higher Education." *Journal of Work-Applied Management* 9, no. 1 (January 1, 2017): 5–17. https://doi.org/10.1108/JWAM-12-2016-0025.

The Guardian. (2019, September 16). "Nike's 'dream crazy' advert starring Colin Kaepernick wins Emmy." The Guardian. https://www.theguardian.com/sport/2019/sep/16/nikes-dream-crazy-advert-starring-colin-kaepernick-wins-emmy

Jensen, J.A., Wanless, L., Hedt, K. and Wayland, E. (2020), "Analyzing Big Data in intercollegiate athletics: an application to athletic donor behavior", Journal of Issues in Intercollegiate Athletics, Vol. 13 No. 1, pp. 292-311.

Kuan, Yun. "An Exploratory Study on Controversy Advertising's Effect on the Advertiser and the Controversial Issue." Portland State University, 2018. https://doi.org/10.15760/honors.561.

Lazer, David, and Jason Radford. "Data Ex Machina: Introduction to Big Data." *Annual Review of Sociology* 43, no. 1 (2017): 19–39. https://doi.org/10.1146/annurev-soc-060116-053457.

Mamo, Yoseph, Yiran Su, and Damon P.S. Andrew. "The Transformative Impact of Big Data Applications in Sport Marketing: Current and Future Directions." *International Journal of Sports Marketing and Sponsorship* 23, no. 3 (January 1, 2021): 594–611. https://doi.org/10.1108/IJSMS-03-2021-0073.

Moreno, J. (2019, December 25). These were the top google searches and trends of 2019. Forbes. https://www.forbes.com/sites/johanmoreno/2019/12/24/these-were-the-top-google-searches–and-trends-of-2019/?sh=26ad38e83089

Nielsen. "Super Bowl LVII Totals More than 113 Million Viewers, Ranks Second Most-Watched Game Ever." 2023. https://www.nielsen.com/news-center/2023/super-bowl-lvii-totals-more-than-113-million-viewers-ranks-second-most-watched-game-ever/.

PDXScholar: The Institutional Repository of Portland State University. Site. (n.d.). https://pdxscholar.library.pdx.ed

Smith, P. by M. D. (2023, January 16). NFL regular-season attendance hit six-year high in 2022. ProFootballTalk. https://profootballtalk.nbcsports.com/2023/01/16/nfl-regular-season-attendance-hit-six-year-high-in-2022/

StarNgage. "Cristiano.Ronaldoo's Instagram Audience Analytics and Demographics." Accessed May 22, 2023. https://starngage.com/app/pk/influencers/cristiano.ronaldoo.

Super Bowl LVII totals more than 113 million viewers, ranks second most-watched game ever. Nielsen. (2023, February 14). https://www.nielsen.com/news-center/2023/super-bowl-lvii-totals-more-than-113-million-viewers-ranks-second-most-watched-game-ever/#:~:text=NEW%20YORK%20%E2%80%93%20February%2014%2C%202023,112.2%20million%20viewers%20on%20FOX.

The role of Big Data in sports marketing. Doxee. (2020, October 6). https://www.doxee.com/blog/digital-marketing/the-role-of-big-data-in-sports-marketing/

Uhlir, Paul F., and Peter Schröder. "Open Data for Global Science." *Open Science Journal* 6 (June 29, 2007). https://doi.org/10.2481/dsj.6.OD36.

Wampler, A. B. (2022, March 17). Researchers develop data-driven tool to optimize sports ticket pricing. Notre Dame News. https://news.nd.edu/news/whats-your-seat-worth/

PART VII

DATA IN PUBLIC RELATIONS, SOCIAL MEDIA, AND ADVERTISING

Chapter Written by Ana'aya McGowan Mozell

Learning Objectives

- Explain the role of data in modern public relations, specifically how it is used for campaign effectiveness, behavioral insights, and message adjustments.
- Gain an understanding of how big data can optimize various aspects of a business, such as resource management, operational

efficiency, and product development, in the context of public relations.

- Discuss the emerging role of Artificial Intelligence in the field of public relations, including its potential benefits and disruptions.

Data is used or generated by nearly everything we do in society. It's no surprise data is a big part of the modern public relations (PR) that we know today. PR businesses use the knowledge of data as their number one priority as a digital marketing strategy since public relations is known as a business of influence and understanding of who, what, and how their audience is connected with them and how they resonate with the brand or business. This strategy gives the brand, organization, or company a well-rounded understanding of what and who their target audience is, their profit numbers, and high and low insights on what could be the next waves of marketing.

But, first, what is public relations?

Public relations is the practice of managing and building a positive public image for a company or organization. The PR experts learn how to create media for their brand for press releases of social media messages that help shape the public's opinion on the brand or company. This helps to increase brand awareness and with the data, – big or small, – it's collected by organizing and analyzing the information. At first glance, public relations and data might not go hand-in-hand together since PR is more focused on social skills than numbers.

But data is a key essential for successful campaigns and better-targeted media outreaches. The data are very ideal key essentials used for proving the effectiveness of PR campaigns, data analysis, behavioral insights, adjusting messaging, and the value of service.

KEY ESSENTIALS

The first data essential is proof of PR campaign effectiveness. Without data, it's not easy to measure the effectiveness and efficiency of a public relation campaign and it's harder to prove that your campaign is influenced by the public decisions to favor your brand or company. Luckily, PR experts are able to record and document the campaigns by spoken words and are able to collect and analyze the data as a present proof of the campaign.

The second data essential is behavioral insights. PR experts gather data by watching and analyzing how a group of people, or the public reacts to what the brand or company is putting out (Lotame, 2022). In a natural way, humans respond more to the environment around them and everything with it. The behavioral data can help you achieve effective ways of making public-approach strategies and making a better image for the company. Understanding how the public reacts and acts towards your company is one of the key functions of public relations.

The third data essential is adjusting messaging. We all know words have a powerful effect on people to influence how someone can communicate or react towards a subject. Inappropriate messaging can damage a brand, an individual, or a company's image in an instance. However, PR experts can craft careful messaging to target a certain group of people and data from trends that can help realign the message the company is sending.

The fourth data essential is the value of service. The quality of the data that is collected to influence the public can help improve your value as a PR expert. The data is a valuable tool in public relations. It shows the growth and success as a brand or company.

BIG DATA

Now, what is the main focus on data? Why is it important for Public Relations?

The focus of data is the relationships revealed by big data. The term "big data" is used to describe data that is hard to manage or too large in masses that can be both unstructured and structured for use (Big Data: What it is and why it matters, n.d.). To save time, PR experts figure out what's important to analyze through the given data. This helps to improve quicker decision making and strategic planning for businesses. The importance of big data is how you use it. By taking the source of data in Twitter, Facebook, or Instagram for example, companies can find their answers in five ways.

The first way is through streamlining resource management (CFI Team, 2023). A great way to optimize a business is streamlining. Companies work better with optimizing effective operations to minimize their cost and profits. It helps businesses get to their highest potential, saving time and money, and minimizing high risks.

The second way is improving operational efficiency. Operational efficiency is important because businesses find ways to reduce costs, waste, improve their productivity, and improve their quality of products and services. It involves keeping track of the company's inputs and outputs as performance indicators as well.

The third way is optimizing product development. Product development is a process of refining and improving a product to make it become more valuable to current consumers and attract new consumers too.

The fourth way is driving new revenue streams and growth opportunities. Revenue streams are the many sources from which a business can earn money. The business can collect sales of goods and services. Revenue drivers is another way for activities, products, services, and marketing that is generated for income to the business as well. This is measured by indicators such as sales volume, price, customer retention, market shares, or growth rate.

The fifth way is enabling smart decision making. By making smart decisions, it requires the brand or company to have as full understanding of the given situation as possible. In most scenarios, this means the company is collecting data from a variety of sources, analyzing what the objective is for the company, and increasing audience engagement to gather evidence on what is working or not.

So why is big data important in public relations?

Data enables the visual of the company's growth and potential success. The collected data can be used and presented as a quality of data to use to influence and improve the public's value as a PR expert. The future of public relations is changing everyday by the new wave of AI.

AI in PR is now becoming a strategic disruption (Hansell, 2022). In recent events, PR experts are discussing the future concepts, benefits, applications, impact and role of artificial intelligence (AI) in public relations.

WRAP UP

Key Takeaways

- Data plays a critical role in public relations, aiding in the measurement of campaign effectiveness, understanding audience behavior, and fine-tuning messaging strategies.
- Big data not only informs PR strategies but also contributes to business optimization by streamlining resource management, improving operational efficiency, and aiding in product development.
- Revenue growth in public relations is increasingly data-driven, with metrics helping to identify new opportunities and evaluate existing revenue streams.

- Artificial Intelligence is becoming a disruptive force in public relations, promising to revolutionize how data is collected and analyzed for strategic decision-making.

Exercises

1. How can data analytics enhance the effectiveness of a PR campaign? Consider a hypothetical or real-world example to illustrate your point.
2. Analyze a recent PR campaign by a well-known company. Using publicly available data, evaluate the campaign's effectiveness in terms of reach, audience engagement, and message clarity. Present your findings in a short report.
3. With the increasing role of AI in public

relations, what ethical considerations should PR professionals keep in mind when utilizing data for campaign strategies?

REFERENCES

Big Data: What it is and why it matters. (n.d.). SAS. https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

CFI Team. (2023, January 15). Streamlining. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/management/streamlining/

Lotame. (2022, November 1). What Are the Methods of Data Collection?. https://www.lotame.com/what-are-the-methods-of-data-collection/

Hansell, G. (2022, July 19). 13 Powerful Ways AI For Public Relations Will Change Over the Next Decade. Demand Generation | SLX.Marketing. https://slx.marketing/ai-for-marketing/13-ways-ai-will-impact-public-relations-over-the-next-decade/

PART VIII

MACHINE LEARNING IN THE DEVELOPMENT OF VIDEO GAMES

Chapter Written by Aboubacar Camara

Learning Objectives

- Understand the three main types of machine learning—Supervised Learning, Unsupervised Learning, and Reinforcement Learning—and their applications in game development.
- Identify modern applications of AI in game development, including Super-Resolution, emulation, cheat detection, and data mining.

- Appreciate the role of AI in the game design industry, including its impact on job roles such as AI Game Programmer and the overall development process.

INTRODUCTION

Usually, when you think about a game being developed you think about the artists, the programmers, and other designers. However, artificial intelligence is increasingly being used in game development. The use of machine learning and deep learning helps developers create new systems and mechanics that feel good for the player. It does this by giving the AI data and seeing how it will process and work out solutions to problems.

AI can play out different scenarios and develop different parts of the game. Artificial intelligence can also be used to increase a game's graphics and visuals through a method called Super-Resolution. There are different types of Super-Resolution methods but the focus for this chapter will be on Super-Resolution Convolutional Neural Networks (SRCNN).

Data mining is a common business practice, but it is used in game development and determines how the game is updated. Before we go into the uses and methods of artificial intelligence, we must first understand the three main types of learning: Supervised Learning, Unsupervised Learning, and Reinforcement Learning.

TYPES OF LEARNING

Designers and programmers may be skilled at developing a product, but they can't predict every problem that will occur. They can't make a solution to every problem, so they use different types of machine learning for artificial intelligence to make a solution that can be implemented into the game. To gain a better understanding here are the 3 types of machine learning:

1. **Unsupervised Learning** is feeding data to a system and applying an algorithm to make observations from the data. In this type of learning the AI will help find patterns in the data to make decisions (Coursera, 2022). This type of learning helps make observations and decisions. Without having an expected outcome the AI is free to make any decisions and solutions to the problems presented. An example of this in research is using player data in an algorithm to understand how players perform in the game. (Drachen, 2009)
2. **Supervised Learning** is feeding data into a system with an expected outcome in mind. In this type of learning the data being fed is made to produce a specific output. In this instance, the AI is being assisted in producing

solutions with the labeled data (Coursera, 2023). More specifically the data you put in will have a corresponding output and you want the AI to learn the relationship between the two. In video games, the goal is to have the AI reproduce player behavior in any situation from the data given to it.

3. **Reinforcement Learning** is when the algorithm or agent can interact with its environment and make a negative or positive reward for the behavior based on the context of the objective (Coursera, 2022). In this type of learning, the AI is learning closest to how the human brain would learn. In game design, this would be good for testing different mechanics and specific missions to see how a player would progress.

Now that the types of learning are understood, one can see that they have many applications in different fields. But, they are also used in other game genres for different purposes. It's crucial to recognize that machine learning's role in game design is not a one-size-fits-all approach. Each type of learning—Unsupervised, Supervised, and Reinforcement—serves unique functions and is best suited for particular challenges within the gaming landscape. Whether it's predicting player behavior, enhancing game mechanics, or understanding player data patterns, machine learning offers innovative solutions to complex issues. Therefore, choosing the appropriate type of machine learning can significantly influence the effectiveness of the game's design and its ultimate success in engaging players. By leveraging these distinct approaches, designers and programmers can create more dynamic, responsive, and captivating gaming experiences.

MODERN APPLICATIONS OF AI IN GAMES

Emulation of Old Games

Artificial intelligence is not only made for developing new games, but it can also be used for redesigning and placing older games onto modern systems. This process is called "emulation" and it is used to put older games that are normally inaccessible to the public onto a more modern system. This has become widely popular on PC systems allowing players to revisit their childhood games that wouldn't otherwise be available. One example of this would be the emulation of the Atari 2600 called Stella. Stella is an emulator used to bring games from old systems such as Atari 2600 and Sega Genesis. The process of emulation is all possible via reinforcement learning. One of the methods used in this learning is the Arcade Learning Environment (Bellemare, 2013). This learning environment was built on the Stella emulator. This environment allowed researchers and others who were interested to add agents and use visual input such as screen pixels to produce an output. The researchers will give the agents specific instructions so the output is a more modern version of the game on modern systems.

Super-Resolution

Another implementation of AI in game development is the use of Super-Resolution. Super-Resolution is the process of increasing the resolution of an image from low to high. It is commonly used in surveillance such as security cameras for facial recognition and in the medical field to produce high-resolution pictures during medical examinations. Super-resolution itself has a couple of methods for how it increases and decreases the overall resolution. One game that uses super-resolution is God of War, which was released in 2016. In 2022, God of War received an update that boosted the game's frame rate and graphics. The result of this is from the super-resolution software FediltyFX Super-Resolution 2.0. This software is designed to upscale the game as you play it. If a game engine runs at 1080p, FediltyFX will boost it to 1440p or 4K resolution (Klotz, 2022).

Super-Resolution Convolutional Neural Network (SRCNN) (Tsang, 2018) uses 3 layers: one for patch extraction, one for non-linear mapping, and the last for reconstruction. Before inputting the image data, the researcher must resize the image to what they want it to be in the end. One approach in super-resolution involves using a multilayered neural network. The initial layer extracts patches from the input and applies filters to represent them. The second layer performs non-linear mapping, preserving the distances between data points while reducing the image's dimensions. The final layer, the reconstruction layer, restores the image after all the processing is complete. The process involves intricate mathematical calculations, but it is only one of many applications of AI.

CHEAT DETECTION

Cheat detection is when the administrator or developers of a game server work to find players who are cheating. The developers will watch how matches are played and pick out those who are suspected of cheating and will look further into them. If the player's activity is recognized as cheating the administrator of the server will either ban or suspend the activity of the player. This application is important because this will help AI in learning what cheats are and helping the game become more efficient and fair for all players.

In every multiplayer game, there are going to be those who you exploit to win matches. Cheaters in video games find a way to win matches with an unfair advantage by exploiting game mechanics and using foreign software. This has caused developers to find out who the cheaters are and punish them for their misconduct. At first, cheat detection was handled by server administrators, but as technology in video games advanced, AI is tasked to detect cheaters and administer bans or suspensions. The AI will take the data from the player on how they perform in the game and determine at some moments whether or not it was a human playing the game. After deciding that it is another AI performing, the anti-cheat will then remove the player from the server. With anti-cheats, it is easier to exploit third-party programs, but with exploits like going underneath the map, it won't be able to always detect that as cheating so it is also up to the development team to either train the AI to see that as cheating, or provide updates to the game to stop exploits. Implementing supervised learning or reinforcement learning by having the AI pick which video input had a cheater in it could be an avenue of training AI. One of the more known anti-cheats is BattlEye, used by most first-person shooter games.

BattlEye is one of the anti-cheat software used in several games such as *Destiny 2*. It is installed as if it was part of the game you are playing. The anti-cheat acts as a shield around the games you play. One of the most known forms of cheating is hacking the game and BattlEye is made to defend against those efforts. BattlEye protects against hackers and runs completely independently without the help of a developer (BattlEye, 2013).

Data Mining

Another application of AI in both video games and business is data mining. Data mining is the process of turning raw data into useful information. Many businesses use this to get an understanding of what their customers want. Whether it be for recommended items or for how to push the business forward, data mining is essential for all businesses to get data on their consumers. This practice is also used by game developers to gain information on how the players behave and how they will play the game. This allows the developers to improve gameplay and make the game more enjoyable. This can also be used by the players who want to know about the plans of the game by leaking information.

AI in the Industry

Having a basic understanding of how AI works is beneficial when going into the game design industry. Going back on the types of learning it is best to understand those three to test how the game works and what outcomes will be produced in the industry, AI is also important for being a game programmer. The position that handles and maintains the AI is called an AI Game Programmer. Their job is to cater the AI to an individual player. Every game studio and publisher will always need programmers to do this job. In games like *Division 2,* you need to use AI to see how a player would typically act or play in a mission or scenario. In other games like *The Last of Us*, the AI of both allies and foes are designed in a way to believe they are human or infected zombies. AI plays a significant role in game development; without it, games would never play or look as good as they are now.

CONCLUSION

AI in the game design industry has a significant impact on the future of games. Video games are usually developed by large teams with each of them having their role. AI can assist them and make their jobs easier. The three types of machine learning are essential in development. Emulation is related to super-resolution which is an example of supervised learning in which the developer wants the image to be upscaled and displayed in higher graphics. Cheat detection makes sure that live service and multiplayer games are played fairly and data mining is helping businesses understand what parts of their game can be improved and changed. AI assistance is the future of game design. The reader should understand why AI is important for most aspects of game design and should go into the game design industry with an understanding of artificial intelligence.

WRAP UP

Key Takeaways

- AI is increasingly becoming a cornerstone in game development, offering innovative solutions to complex problems like game mechanics and player behavior prediction.
- The three types of machine learning—Supervised, Unsupervised, and Reinforcement Learning—serve distinct functions and are applicable to different challenges within the gaming landscape.
- Modern applications of AI in games extend beyond gameplay mechanics to include emulation of old games, enhancing graphics through Super-Resolution, cheat detection, and data mining for game improvement.

- A basic understanding of AI and machine learning is becoming crucial for roles in the game design industry, including specialized positions like AI Game Programmer.

Exercises

1. How do the three types of machine learning differ in their applications within game development? Provide examples for each.
2. Choose a modern game that you believe utilizes AI in its development. Investigate and present how AI has been used to enhance the game, whether it be through game mechanics, graphics, or player interaction.
3. Considering the role of AI in cheat detection, what are the ethical implications of using machine learning to monitor player behavior?

Should there be limits to how this technology is used?

REFERENCES

AI and Games (Director). (2023, April 18). *How Machine Learning is Transforming the Video Games Industry | AI 101*. https://www.youtube.com/watch?v=dm_yY-hddvE

BattlEye – The Anti-Cheat Gold Standard » About. (n.d.). Retrieved April 18, 2023, from https://www.battleye.com/about/

Bellemare, M. G., Naddaf, Y., Veness, J., & Bowling, M. (2013). The Arcade Learning Environment: An Evaluation Platform for General Agents. *Journal of Artificial Intelligence Research*, *47*, 253–279. https://doi.org/10.1613/jair.3912

Coursera. 3 Types of Machine Learning You Should Know. (2022, September 14). https://www.coursera.org/articles/types-of-machine-learning

Justesen, N., Bontrager, P., Togelius, J., & Risi, S. (2020). Deep Learning for Video Game Playing. *IEEE Transactions on Games*, *12*(1), 1–20. https://doi.org/10.1109/TG.2019.2896986

Klotz, Aaron. (2022, June 1). *AMD's FSR 2.0 Delivers Again in God of War*. Tom's Hardware. https://www.tomshardware.com/news/fsr-2-0-god-of-war-good-results

Spectrum, C. G. (n.d.). *AI Game Programmer Job Description, Salary, Skills & Software*. Retrieved April 25, 2023, from https://www.cgspectrum.com/career-pathways/ai-programmer

Tsang, S.-H. (2022, June 24). Review: SRCNN (Super Resolution). *Coinmonks*. https://medium.com/coinmonks/review-srcnn-super-resolution-3cb3a4f67a7c

PART IX
THE USE OF MATCHMAKING DATA FOR COMPETITIVE ONLINE MULTIPLAYER GAMING

Chapter Written by Glenndale Bartolome

Learning Objectives

- Understand the complexities and challenges of online multiplayer matchmaking algorithms, including methods like the Elo ranking system and TrueSkill.

- Identify the different metrics and data used in ranking systems to assess player skill and ensure fair gameplay.
- Appreciate the ethical and social implications of data collection in online multiplayer games, including issues related to player satisfaction and game balance.

INTRODUCTION

Video games have come a long way. We can play with other people from across the world, something that seemed unimaginable in the 80's & 90's. This phenomenon, called online multiplayer, has been a big thing in the gaming community, and a lot of games use algorithms to decide which players fight each other. Games like first-person shooters (e.g., Halo, Call of Duty, Apex Legends), sports games (e.g., Fifa, Madden, etc.), and fighting games are generally expected to have a proper functioning online component at launch if they want to succeed at all. A significant amount of data is used to make sure game features are set up functionally and fairly. For example, this data might include things like ping and frame rate or a player's win/loss ratio. This data is run through algorithms that help define the skill level and internet connection quality of each player. They primarily use a system similar to the Elo ranking system of Chess, which uses two primary equations. The first one measures the probability for one player to win over another, and the second one takes that result and uses it to determine that player's rank.

Elo Rating System Equations (Véron et al., 2014, sec. 2.2):

- $Ea = \dfrac{1}{1 + 10^{(Rb - Ra)/400)}}$
- $Ra_{new} = Ra_{old} + K * (Sa - Ea)$

While this system works, it has created challenges for some players. Depending on the game, this matching system can either match really good and famous players with really bad ones and drag their ranking down, or in some games, these content creators are actively looking for lower ranked players so they can dominate the competition and gain lots of viewers. Though they aren't a majority, these big names also advertise the games they play and broadcast to the world. This means developers need to somehow cater to both sides and find a way to keep the game from getting imbalanced. They do this through the use of the data of all players, which is shown in various multiplayer games. But this data collection also poses certain problems regarding the secrecy of development, and just how some people get access to the data.

THE MATCHMAKING PROCESS

To get an understanding of the problem, we first need to understand how matchmaking works. However, the way it works differs from developer to developer, and game to game. So let's take a look at some case studies to see how different games handle matchmaking. We've introduced Elo, which a lot of games use, but here's some alternative examples. First, let's look at TrueSkill, the matchmaking system for games like Halo. According to one of the creators, "[the] rating is a Gaussian distribution which starts from N(25,). μ is the average skill of a player, and σ measures how likely this is true. [The] real skill of a player is between $\mu \pm 2\sigma$ with 95% confidence." (Lee, 2012, para. 4). These are applied through the multiple algorithms that manipulate the rating depending on the type of match, determining how players would do against each other. Now, this is definitely a workable system. It's been the backbone of Halo matchmaking since its inception, and has made its way into countless other games. The issue is that the data for it is locked to a set of systems, and that leads to some matches that have some bad experiences.

Another challenge with this approach is that it has difficulty balancing a predefined level gap which defines match balance and the amount of players of an appropriate level of skill. This has caused a rift in the developers, common players, and some content creators and professionals, who think that their skill rating doesn't match with their actual skill. One example of a pro player's perspective is the player Eric "Snip3down" Wrona, who stated in an interview recorded by Ethan Davidson that "I'm one of the best players in this game and I'm losing 70 percent of my games, how is this possible?" (Davison, 2022, para. 9, citing Wrona). This was a complaint to 343 Industries about how matchmaking in the game "Halo 5: Guardians" caused him to feel as if he was set up not to enjoy the content. The mindset seen in content creators like him is one where, rather than wanting to be matched up with people's skill level or thereabout, they instead desire matches in which they can generate popular content by winning matches, fighting against casual players. But they aren't a majority.

One thing that Thore Graepel and Ralf Herbrich state to the Game Developer Magazine of Microsoft emphasizes the importance of "the purpose of the game and the behavior of the rating system [being] aligned: people striving for high ratings should be forced to play in accordance with the spirit of the game. Taking the margin into account by which a game was won can be very misleading." (Graepel & Herbrich, 2006, p. 3). To them, the clean sweeps that Wrona and streamers like him desire are detrimental to the other players, some of them wanting to genuinely improve, not get bullied by pro players for views and revenue. Still, Wrona's view persists, and even some developers are agreeing. One example is Max Hoberman, designer of ranking systems in Halo 2 & 3, who stated that "perfectly balanced games…were often the most stressful." (Davison, 2022, para. 27). With this big divide, a solution seems unclear, though the game Farmville may hold a temporary solution to this issue.

ANALYZING OTHER APPROACHES

In Davidson's article, he described at length a matchmaking system based on player engagement that he reported came from Zhengxing Chen, a researcher at Facebook. In it, he mentioned the amount of additional data, such as the time it takes for them to put down a game, and alters the next match so that the player is more likely to want to keep playing. This was tested by Farmville, according to a researcher named Naomi Clark whom Davidson cited. It seemed to work too, although Farmville is also single player. However, it could work for multiplayer games as well, taking into account things like the ratio of wins and losses, game duration, number of games played, and whatever data is exclusive to a certain type of game mode. Trueskill, for example, has sets of rules for 16-player free for all games, as well as games that have either two teams total, or four teams total.

Figure 1: List of Rules in the Trueskill rating system (Lee, 2012, tbl. 1)

Rule	**Matches**
16P free-for-all	3
8P free-for-all	3
4P free-for-all	5
2P free-for-all	12
2:2:2:2	10
4:4:4:4	20
4:4	46
8:8	91

Multiplayer matching could anticipate complaints and address them appropriately by using player data. But it could also ruin a game's objective and/or subjective fairness, which is arguably more important, as stated by Herbrich and Graepel, in their study where they stated that, "Matchmaking should be based primarily on skill and be otherwise not under the influence of the gamers. Ranked re-matches should be disallowed or limited to one to avoid the risk of collusion." (Graepel & Herbrich, 2006, p. 6) This collusion can ruin the subjective fairness that lower-level players will have seeing one higher-level player gain a major boost, just because they've been losing too often, and as shown in Figure 2 by Véron and the others, the greater the skill gap between players, the more likely players are going to quit the match. Along with this is the concept of “smurfing”. This is when experienced players start up new accounts and pretend to be newer players, and play against actual new players and casuals. The result, according to a group of researchers looking into another MOBA game “Heroes of Newerth”, is: “thus winning easily, but ruining the playing experience for inexperienced, and often new, players in the process (and cutting into the future profit for the company, as well).” (Caplar et al., 2013, p. 2). It goes to show that these competitive players are going to conflict with the casual audience just trying to have fun, or with people trying to grow their skills independently. So let’s consider the research of Neven Caplar and teams’ studies and see what they think.

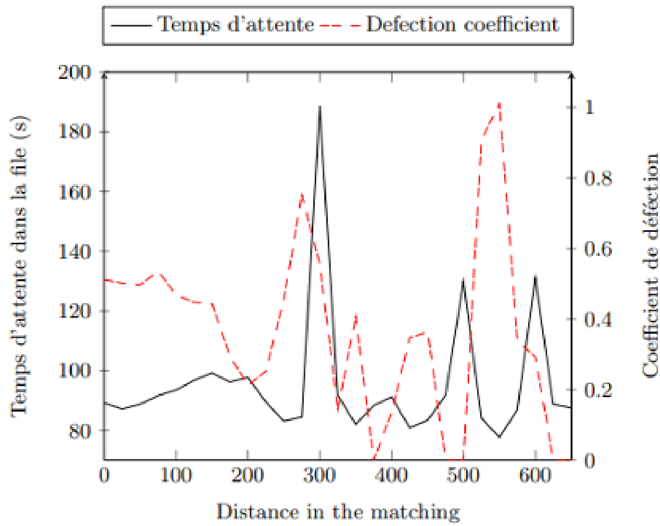


Figure 2: Distance between players skill levels and frequency of players quitting with waiting times thrown in as a control (Véron et al., 2014, fig. 2)

In researching how to deal with this smurf issue, they first sought to gather a dataset of player ratings for their case game, which was done by taking the whole player ladder, which has been made available on the website www.honedge.com. This site has since been abandoned. They then took the statistics of several thousand players and did some math to study the player's ranking. Player ranking is decided via Elo rating, and in it they discovered its limitations. Those of note include "rating inflation, and freezing of top rankings (by players who stop playing once they have reached top positions, i.e., no rating deflation over time)." (Caplar et al., 2013, p. 2). They also looked into its matchmaking algorithm, and made some interesting comments. According to the researchers, the developers of the game, Garena and S2 Games, posted a patch that supposedly "addressed [the] recognized problem of 'smurfs'." (Caplar et al., 2013, p. 2).

They also touched on the possibility of using neural networking. To paraphrase, they cited another scientific study that proposes using neural networks to evaluate the skills of players and maximize their perceived fun factor, as well as predict complex team scenarios where they might not be even in terms of members. This complex use of neural networks could possibly be the solution for this big issue. They mentioned how matchmaking shouldn't solely have to be based on player skill. It could be easier to base it off of network connection, if we borrow from their example. This is actually something that Davison cited Chen using, in a phone interview where "[he] confirmed the growing complexity of matchmaking techniques: 'Previously, they only looked at your win-loss history … and tried to develop one scalar score [like Elo or MMR] for you to summarize your skill. But as time goes on, I can see that there's work using neural networks to summarize your skills in multiple aspects, not just one single score, and trying to use more history, more information to estimate your skills in different areas.'" (Davison, 2022, para. 21). This could be done with the amount of metrics that rating systems already gather, such as win/loss ratio, experience & currency per minute, how much time or real-world money a player has spent on the game, the length of time the game lasts, and so much more.

The rest of their experiment demonstrated these metrics in use and how they affected the matchmaking rank (MMR). In section 5, the results of their large case study were unveiled. The first subsection demonstrates how the number of games played affected ranking, demonstrated in the graph seen below.

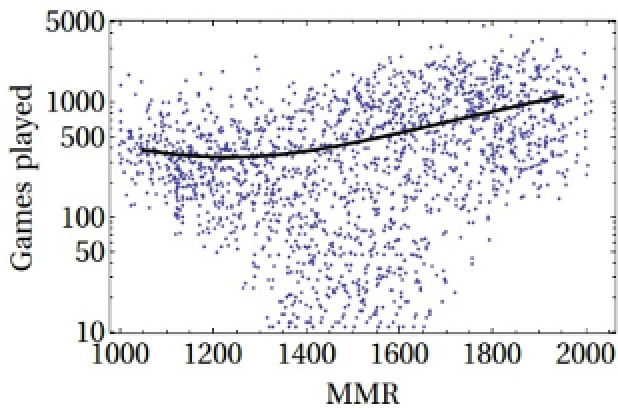


Figure 3: Number of Games Played as Function of MMR (Caplar et al., 2013, fig. 3)

This graph has a correlation, sure, but it contains some anomalies, namely, involving the trend line having no feasible way to match with the results due to an extreme amount of variance in people's MMR and the number of games played. The next subsections involved ratios. including the number of wins and the number of losses, and the ratio of kills, assists, and deaths. These revealed a rather obvious common trend of the higher one's rank reflecting a higher number of kills and assists. Afterwards is gold (currency) and experience a player gains in a minute. Experience is a number type that determines the overall effectiveness of a character. High experience means more levels, which means a character is stronger. The result of it is that more skilled players are able to get these things much easier, and experience can be picked up by anyone, meaning that matches would likely match those of similar rank together because they can better coordinate things so that if a player is nearby, they can both gain experience from a person's kill.

Afterwards is game length, action rate, rate of spawning wards (an item that allows map visibility), denying players of killing your creeps (NPCs that help attack bases, minions in other games), and a player account's age. Game length had two graphs, one which showed that the probability for a match to end at a certain time decreased the higher the time was, generally ending at the 20 or 40 minute mark, 40 minutes for the full match, and 20 minutes if the game was called off early. The other showed that games were often shorter in higher ranks on average.

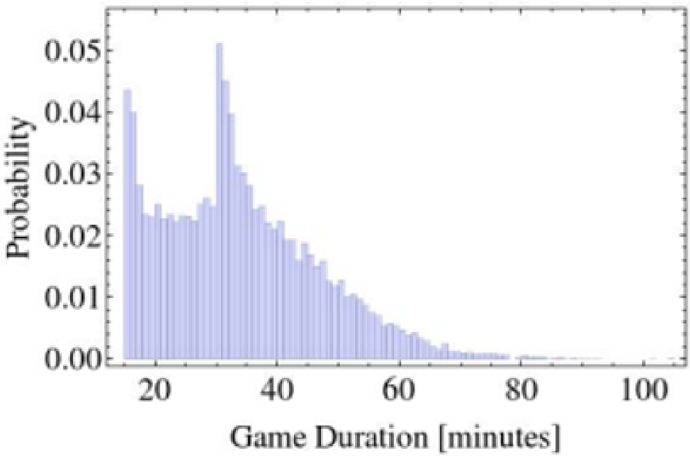


Figure 4. Distribution of games duration (Caplar et al., 2013, fig. 4)

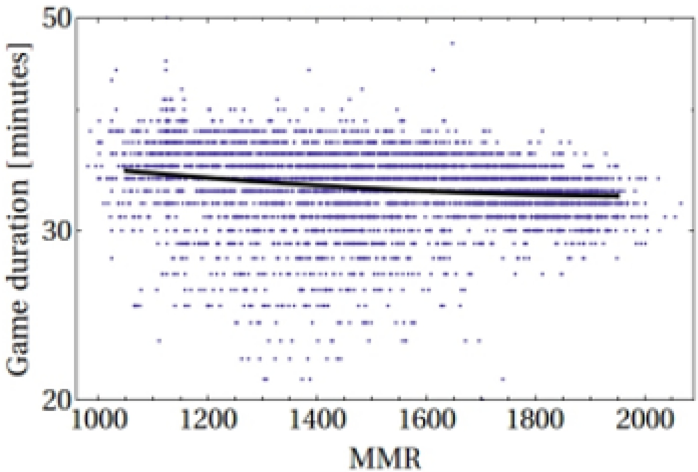


Figure 5: Average game length as function of MMR (Caplar et al., 2013, fig.5 (6)

The 20 minute mark ends and the quitting system seems to help deal with griefers, players who intentionally sabotage their team, as four out of five team members have to approve to quit the match.

Next is action rate, the rate players performed actions, which increased with barely a curve in the trend line. It seems to conclude the same thing about match formation that k/d/a ratios and win/loss ratios do, encouraging that similar ranking. Warding rate is next, and this one had a sharp increase in the beginning, but after a certain amount of use, it curved over into and slowed down increasing. The same for the number of denials of creeps, which seems to imply that for lower ranked players, these are viable uses of time and resources, something that experienced players don't need to use as often, as they know the counters. Finally, the age of a player's account, which has a heat map, but depicts a result similar to the third figure Caplar and the others created. In this study, they concluded that this data collection with some error, does at least manage to make a fair assessment of people's performance. But it's too slow assigning them to said skill groups. So is there still a way to speed things up?

Maybe there is, and that could be a Peer-to-Peer (P2P) system of matchmaking called the SelfAid, as proposed by Michał Boron, Jerzy Brzezinski, and Anna Kobusinska. They state in their article that the "...presented solution allows a player to quickly connect to others, provided that no failures occur. In this case, accessing a service algorithm is only a matter of issuing one request to announcement DHT and then one request to the process." (Boroń et al., 2020, sec. 7). The Distributive Hash Table is obtained through a service algorithm which contains the necessary data to help match players into the place they want, all without the need of a server.

CONCLUSION

In conclusion, matchmaking in multiplayer games has evolved into a complex and data rich process. Still other variables beyond those discussed could be taken into consideration, such as where each player lives. As technology progresses and people's opinions change, there may be a time in which every person can eventually be satisfied with the game that they are about to play.

WRAP UP

Key Takeaways

- Matchmaking in online multiplayer games has evolved into a complex process that relies heavily on algorithms and data to ensure fair and engaging gameplay.
- Systems like Elo and TrueSkill are commonly used, but they come with challenges such as inaccurate skill assessment and the potential for imbalanced matchups, affecting both regular and professional players.
- Data collection in online multiplayer games is not without ethical considerations; it can affect the secrecy of development and raise questions about who has access to the data.
- Emerging solutions like neural networks and

Peer-to-Peer (P2P) systems such as SelfAid could offer more dynamic and adaptive approaches to matchmaking, although they bring their own set of challenges and considerations.

Exercises

1. How do different ranking systems like Elo and TrueSkill address the challenges of creating balanced matches? What are their limitations?
2. Research and present an example of an online multiplayer game that has faced criticism for its matchmaking system. Discuss the issues raised and any proposed or implemented solutions.
3. Considering the ethical implications of data collection in online multiplayer games, what

are the potential risks and benefits? Should there be limitations on what data is collected and how it is used?

REFERENCES

Boroń, M., Brzeziński, J., & Kobusińska, A. (2020). P2P matchmaking solution for online games. *Peer-to-Peer Networking and Applications*, *13*(1), 137–150. https://doi.org/10.1007/s12083-019-00725-3

Caplar, N., Suznjevic, M., & Matijasevic, M. (2013). *Analysis of player's in-game performance vs rating: Case study of Heroes of Newerth* (arXiv:1305.5189). arXiv. http://arxiv.org/abs/1305.5189

Davison, E. (2022, May 27). *What is skill-based matchmaking and why do streamers hate it? – The Washington Post*. The Washington Post. https://www.washingtonpost.com/video-games/2022/05/27/skill-based-matchmaking/

Graepel, T., & Herbrich, R. (2006). *Ranking and Matching: How to rate players' skills for fun and competitive gaming*. 20.

Lee, H. (2012, January 12). *TrueSkill—Trueskill 0.4.5 documentation*. https://trueskill.org/

Véron, M., Marin, O., & Monnet, S. (2014). Matchmaking in multi-player on-line games: Studying user traces to improve the user experience. *Proceedings of Network and Operating System Support on Digital Audio and Video Workshop*, 7–12. https://doi.org/10.1145/2578260.2578265

PART X
VIDEO GAMES, MICROTRANSACTIONS, AND DATA

Chapter Written by Keshauni Johnson

Learning Objectives

- Understand the historical development of microtransactions in video games, from their early roots to their current ubiquitous presence in both mobile and triple-A titles.
- Explain the ethical and psychological mechanisms at play in the use of microtransactions, including the concept of the sunken cost fallacy and the role of data analytics in shaping purchasing options.

- Gain insights into the economic impact of microtransactions on the video game industry, including revenue generation and market trends.

INTRODUCTION

The video game industry has evolved rapidly since the days of the Atari. With the rise of online gaming, the emergence of microtransactions grew along with the industry. Microtransactions have become an increasingly common feature of video games. Microtransactions refer to small purchases made within a game, often for virtual goods or premium features. They can be found in a wide range of games, from mobile games to triple-A titles, and have proven to be a lucrative source of revenue for game developers and publishers. However, microtransactions have also generated controversy among gamers and industry observers. When microtransactions were small, cheap, one-off cosmetic items, they weren't looked at in the same controversial way as they are now. In this chapter, I examine the phenomenon of microtransactions in video games, exploring their history, how they use data, and the data that reveals their impact on the industry and the players.

HOW COMMON ARE MICROTRANSACTIONS?

Microtransactions are most likely to be found in mobile games and games that are free to play such as Overwatch 2, Fortnite, Genshin Impact, and The Sims 4 to name a few. All of these games are free to download and jump into from the start, however, as you play the game, you'll quickly find they have significant content that requires payment. One example of this is a battle pass system, for which you spend a certain amount of money each month or season to unlock a set amount of content like character skins or weapon wraps through playing and leveling up the pass, similar to Overwatch 2 and Fortnite. Alternatively, as in the case of The Sims 4, you get access to the base game but content like owning pets, getting to start a farm, having your Sims go to university, or even experiencing the seasons like fall and winter is all additional content that must be bought separately, which can quickly add up. Microtransactions can range from items like convenient time skips that speed up or finish production on items in games such as Cookie Run Kingdom to items that boost gameplay to make it easier in games such as Candy Crush.

A BRIEF HISTORY

Microtransactions have their roots in the early days of online gaming. Online multiplayer games such as Everquest and Ultima Online, both of which came out in the late 90's to the early 2000s, offered players the ability to purchase virtual items and currency using real-world money and are some of the earliest examples of microtransactions in gaming. The reaction to them in those early days wasn't nearly as divisive as it is now. In the early 2000s, microtransactions in video games first cropped up within games after DLC (Downloadable Content) rose to prevalence within the industry along with the rise of the modern internet. Though DLC existed before the 2000s in a much lesser form, the concept opened doors for the medium as a possibility to add content to a game post-release to potentially fix bugs or add on to an already finished game as bonus content such as simple cosmetics that were given for free. However, when microtransactions became more common around 2006, their predatory nature raised questions that included how they may potentially be predatory and abuse the psychological mechanism of fear of missing out (FOMO).

Often, the DLC that is created is backed up by data gathered about their players or outside data from trends throughout the industry. This data guides decisions on how they should implement the microtransaction system effectively in their games in ways that will yield the most profits. For example, at the most basic level, let's say you're making a game that is an online multiplayer game and you want to make a profit. If you had the option to choose between earning a flat 60 dollars now or 100 dollars or more over time per player that invests into your game, which would you choose? This is something that companies think about when deciding on how to monetize their games. Now you'll see microtransactions often seen in live service games that continually update or in a plethora of mobile games on e-storefronts like Apple's App Store and Android's Google Play store. These games can continually collect data about which items are yielding the most sales, allowing them to tailor future offerings in ways that are most likely to appeal to to a wide audience and trigger further purchases.

HOW QUICKLY CAN IT ADD UP?

The rise in popularity of these types of free-to-play games raises significant ethical questions, including about whether or not these practices are predatory. For example, many have argued that these games rely on the sunken cost fallacy. This fallacy describes a situation where players that have spent hours and hours in a game and then no longer want to play or no longer have fun playing feel as if they are unable to simply quit playing and just walk away from the game. However, the time and money that was spent on the game can't be recovered, making it appear worthwhile to continue playing a game into which one has "invested" these resources . When you have spent hundreds to thousands of dollars in a game through buying in-game currency or items, it makes it a lot harder to walk away. Players can feel that they've made too much of an investment in the game to simply drop it.

Another point of debate is how these games collect far more money over time from the average user than a company would for just a simple 60-70 dollar one-time purchase of a triple-A title. For example, if you were to buy a typical third-person shooter at full price for 60 dollars, you would have access to all the content and online features. However, in games like

Fortnite, the price to download and start playing is zero, however, if you buy the season's current battle pass for $9.99, play through it, but then buy some V-Bucks, which is the in-game currency, you could end up spending an additional 10-20 dollars. With new skins and other items rotating in the shop every day, you could say, "Well it's just one more skin, I like this one," and buy that. Then a collaboration, or collab, comes along that interests you and you buy something from that, spending another 20 dollars to get the collab set. The longer you play, the more you're likely to buy smaller things like emotes based on popular dances and wraps that change the outer look of a player's weapon like in Fortnite or Overwatch 2, for example.

Players, especially younger ones, can feel compelled to buy cosmetics like costumes for their character to wear for several reasons ranging from social pressure to not wanting to appear 'poor' or lower skilled. According to a study done in the UK, children often see the type of skin you have as a status symbol within the community meaning they were more likely to ask their parents to buy them in-game currency so they could obtain better cosmetic items (Wood, 2019). These can quickly add up, and in a live service game like Fortnite that constantly gets updates, the average player soon finds themselves spending way more than 60 dollars on Fortnite, which is far more than they would have just by buying the one triple-A title at a flat price.

In the year 2018, players spent an average of $84.67 on in-game purchases (DemandSage, 2023). Comparatively, in the year 2020, an average of $102.42 was spent by Fortnite players (Statista, 2022). Companies and developers are aware of this, as these games were made with these models in mind from day one. Games like Candy Crush have an older demographic where 50% of the Candy Crush players are aged between 20 and 40 years old (EpicWinApp, 2023). These games make a ton of money and microtransactions were implemented within the game from the start. Candy Crush alone earned $77 million the year after its release in 2012. They earned $1.13 billion two years later (EpicWinApp, 2023).

IMPACT OF MICROTRANSACTIONS

The impact of microtransactions on the video game industry has been significant. For game developers and publishers, microtransactions represent a lucrative source of revenue. In 2020, the global video game market was estimated to be worth $159.3 billion, with microtransactions accounting for a substantial portion of that figure (Global Games Market Report, 2022). For some games, microtransactions can generate millions of dollars in revenue each year. Only 5 to 20% of game communities take part in microtransactions, and the amounts they spend vary (Investopedia, 2022). Developers and publishers will look at these numbers and trends within the market to see what will potentially make them the most amount of money. With the heavy dominance of mobile gaming, thanks to the ease and accessibility of smartphones, microtransactions in games aren't going anywhere anytime soon. Live service free-to-play games like Fortnite have set a prominent trend in the industry, showing how this strategy of designing, building, and marketing a game around this model of monetization is extremely lucrative. The investment over time operates similarly to subscription services for streaming platforms and the long game for these microtransaction-based games has shown the power perspective.

Some of the ethical issues players often talk or debate about when it comes down to microtransactions in games are implementation and effect. Microtransactions aren't inherently a bad thing, but often it is the way in which it is being used within a game can be problematic. Examples of this can be games that can limit your playtime unless you pay to remove the limit, games with a battle pass that make you commit to playing to finish the pass or you risk losing out on the money you spent, and limited content that won't ever come back so that you have to play or you risk losing out (Neely, 2021).

CONCLUSION

Microtransactions have become a common feature of video games, generating significant revenue for game developers and publishers. When a couple of dollars here and there from a player base of thousands and sometimes millions over the span of years is compared to the initial launch year of a 60-dollar triple-A game, the optimal choice for most developers looking for a big profit is clear. While incredibly profitable now, only time can say how long they will stick around within the gaming industry. However, it's undeniable how profitable the model is and has been over the years and how prevalent the trend is. On the other hand, players haven't been the most receptive to the growing dominance of microtransactions in the industry. Players have been vocal with their dislike of the oversaturation of microtransactions in games and how greedy some of the uses of microtransactions within these games appear to be. However, when it comes down to the numbers, the industry of microtransactions has only grown and is projected to grow even more in the years to come.

WRAP UP

Key Takeaways

- Microtransactions have evolved from being small, cosmetic items to intricate systems that are integral to many games, especially in free-to-play models, often leveraging psychological mechanisms like FOMO (Fear of Missing Out).
- These in-game purchases are highly profitable for game developers and have a considerable impact on the video game industry, sometimes surpassing the revenue generated from initial game sales.
- Ethical questions arise from the use of microtransactions, especially when considering the sunken cost fallacy, which

keeps players investing time and money even when they no longer enjoy the game.

- Data analytics play a crucial role in optimizing microtransactions for profitability, allowing developers to tailor offerings based on which items are yielding the most sales.

Exercises

- Discuss the ethical implications of microtransactions. Do you think they are inherently predatory, or can they be implemented in a way that is fair to players?
- Analyze the microtransaction systems in two different games—one mobile and one triple-A title. Compare their strategies, ethical considerations, and how they impact the player's experience.

- Given the economic benefits for developers, do you think microtransactions are a sustainable model for the future of gaming? Why or why not?

REFERENCES

Stephenson, J. (2023, March 10). *2023 Candy Crush statistics for the avid fan*. epicwin. https://www.epicwinapp.com/candy-crush-statistics/

Ruby, D., & About The Author Daniel Ruby Content writer with 10+ years of experience. I write across a range of subjects. (2023, February 7). *Fortnite statistics for 2023 (users, Revenue & Devices)*. Demand Sage. https://www.demandsage.com/fortnite-statistics/

ReportLinker. (2022, March 22). *Online microtransaction global market report 2022*. GlobeNewswire NewsRoom. https://www.globenewswire.com/news-release/2022/03/22/2407893/0/en/Online-Microtransaction-Global-Market-Report-2022.html

Colagrossi, M. (2022, February 8). *How microtransactions impact the economics of gaming*. Investopedia. https://www.investopedia.com/articles/investing/022216/how-microtransactions-are-evolving-economics-gaming.asp#:~:text=Game%20developers%20have%20learned%20to,free%2Dto%2Dplay%20games.

Clement, J. (n.d.). *Topic: Gaming monetization*. Statista. https://www.statista.com/topics/3436/gaming-monetization/#topicOverview

Wood, C. (n.d.). *Kids who play fortnite say they get bullied and shamed if they can't afford paid skins, according to a damning report on gaming habits*. Business Insider. https://www.businessinsider.com/kids-feel-poor-if-they-dont-buy-custom-fortnite-skins-2019-10

Neely, E. L. (2021, April 23). *Loot boxes, microtransactions and freemium games: The ethical considerations*. PlayLab! Magazine. https://www.tuni.fi/playlab/loot-boxes-microtransactions-and-freemium-games-the-ethical-considerations/

PART XI

ARTIFICIAL INTELLIGENCE IN STRATEGIC COMMUNICATION

Learning Objectives

- Analyze and evaluate the impact of artificial intelligence on various aspects of communication and collaboration.
- Develop strategies for the ethical and effective implementation of AI technologies in communication and collaboration, considering factors such as privacy, security, potential biases, and the need for human-centered design.

- Apply critical thinking and problem-solving skills to real-world scenarios involving AI-enhanced communication and collaboration, demonstrating the ability to propose innovative solutions, analyze potential challenges, and assess the societal implications of AI-driven interactions.

INTRODUCTION

Since ChatGPT launched in November of 2022, it has taken the professional and academic world by storm. Professionals have quickly adopted generative artificial intelligence (GenAI or sometimes GAI) into their workflows and schools of all levels scrambled to understand how how such the widespread availability of these writing tools impacts their classroom practices. Who is using GenAI already? Marketers and social media managers have quickly embraced the new tools. Medical professionals are experimenting with using GenAI to help them write reports that are more readable and accessible for patients. GenAI can also translate quickly and easily, with direct voice-to-voice live translation announced by OpenAI in May of 2024.

In higher education, some professors are focused on banning this technology entirely, opting to move backward toward in classroom assessment based on hand-written essay prompts and tests. Heated discussions rage over whether or not tools are able to detect the use of GenAI in student work and whether or not its even ethical to use such tools. Others are revising all of their assignments to accommodate or incorporate GenAI technology. This group has largely made the argument that if professionals are using GenAI in their daily workflows, then schools need to be teaching those skills, framed with clear ethical guidelines.

GenAI hype surrounds us on a daily basis, but so does substantial fear and anxiety. Many worry that such tools will continue to erode critical thinking skills, or remove something that is essentially "human" from the creative process. Others believe that because GenAI tools are trained on the writing and artwork of humans, all use of such tools is a form of intellectual and creative theft. Will the technology continue to improve and eventually achieve sentience?

These are the questions of which we all collectively must strive to make sense. This chapter aims to give an overview of some of these major issues while also demonstrating how to use a variety of GenAI-based tools that might increase the productivity and creativity of professionals. For each topic, I offer an overview, a tool demonstration, suggested readings, and suggested assignments, as ways to help develop deeper understanding of GenAI and how it might be used professionally and ethically.

For transparency reasons, I will also note that I used GenAI extensively in creating this chapter. The header image for each chapter was created with MidJourney. The videos were edited with a variety of GenAI tools that are also demonstrated in these videos. It was also used for some text editing purposes, though all content was originally written by me.

AI AND SOCIETY

[Unedited image created by AI. Errors maintained for demonstration purposes.]

Overview

This section covers the societal implications of AI, emphasizing the importance of ethics and the current uses of ChatGPT and GenAI. It covers notable headlines, such as the controversy with Sports Illustrated's use of AI-generated articles and fake profiles, and explores AI's impact on productivity, job displacement, and creative processes. Additionally, it introduces the concept of prompt engineering and its significance in optimizing AI outputs, and provides further reading suggestions for those interested in deepening their understanding of AI's effects on various aspects of society. Finally, it demonstrates how to use text-based generation tools such as ChatGPT and features examples of using the voice interface.

Videos

AI & Society

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=313#oembed-1

Generative Text Part 1

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=313#oembed-2

Socrates Bot

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=313#oembed-3

Generative Text Part 2

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=313#oembed-4

Mock Job Interview

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=313#oembed-5

Suggested Readings:

1. *AI Training*
2. ChatGPT Cheatsheet
3. Anthropic Prompt Library

Suggested Assignment:

Assignment Description: Exploring Prompt Engineering with LLMs

Overview:

This assignment aims to introduce you to the concept of prompt engineering through hands-on experience with at least two different Large Language Models (LLMs). You will explore how different prompts can influence the responses of these models and develop a deeper understanding of how to effectively communicate with AI technologies.

Objectives:

1. Gain practical experience in designing prompts for LLMs.
2. Understand the impact of prompt design on the responses generated by AI.
3. Compare and contrast the effectiveness of different LLMs in understanding and responding to prompts.

Instructions:

1. **Select Two LLMs:** Choose two LLMs from the following list: OpenAI's ChatGPT, Anthropic's Claude or any others from this list.
2. **Develop Prompts:** Create at least three unique prompts that you can try on each LLM. Refer to the ChatGPT cheat sheet and the Anthropic Prompt library in the readings for help in crafting prompts. These prompts should be designed to test the model's ability to understand and generate relevant and coherent responses. Consider varying the complexity and specificity of your prompts. Try iterating your prompts

to get better reponses as you go. Consider prompts that simulate real-world scenarios where AI might be used in communication and collaboration. This can involve customer service interactions, team meetings, or negotiations where AI tools provide support or automation.

3. **Document Responses:** Record the responses from each model to your prompts. For ChatGPT you can share a link to the chat. For others you may need to take screenshots or copy and paste the text. Note any significant differences in how the models handle the same prompt.
4. **Analysis:** Write a 500-word analysis comparing the performance of the two LLMs. Discuss which model performed better and hypothesize why certain prompts worked well or poorly with each model.

Submission Requirements:

- A document containing your prompts, the responses from the LLMs, and your analysis.
- Format your submission as a PDF.
- Include screenshots or direct text outputs from your interactions with the LLMs.

Rubric for Prompt Engineering Assignment

Criteria	**Excellent (90-100%)**	**Good (80-89%)**	**Satisfactory (70-79%)**	**Needs Improveme (<70%)**
Prompt Creativity	Prompts are highly creative and effectively test different capabilities of the LLMs.	Prompts are creative and have a clear purpose.	Prompts are somewhat repetitive and lack clear objectives.	Prompts are not effective are too simplistic.
Quality of Analysis	Provides a deep, insightful comparison of the LLMs with detailed explanations supported by specific examples from the responses.	Analysis is well-reasoned with some specific examples.	Analysis covers basic observations without much detail.	Lacks depth critical analy of the LLM responses.
Clarity and Organization	Submission is exceptionally well-organized, with clear documentation of prompts and responses; analysis is coherent and logically structured.	Well-organized submission and analysis with minor clarity issues.	Organization is adequate, but some parts may be confusing or poorly structured.	Poor organization and lack of clear structu in documentati and analysis.
Adherence to Submission Guidelines	Fully adheres to all submission requirements and guidelines.	Mostly adheres with minor deviations from guidelines.	Meets the basic requirements but misses some elements.	Fails to meet multiple submission guidelines.

ENHANCING CREATIVITY AND PRODUCTIVITY WITH AI

[Unedited image created by AI. Errors maintained for demonstration purposes.]

Overview

This section explores the impact of generative AI on productivity and creativity, with an emphasis on how GenAI can assist in various stages of the writing process, such as generating ideas, providing inspiration, and ensuring consistency in style. Additionally, it addresses the complexities of GenAI's influence on writing, including legal concerns, questions of originality, and the balance between using AI as a tool and preserving the integrity of human creativity and learning. It explores the following tools: Arc browser, Superhuman email, Descript, and AlteredAI.

Videos

Productivity and Creativity

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=315#oembed-1

Arc Browser

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=315#oembed-2

Superhuman Email

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=315#oembed-3

Descript

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=315#oembed-4

Funny story – my AI process made some slight alterations to the supposedly unedited demo video in the demonstration above. I'm sharing the full unedited and edited copies below so you can see the difference clearly.

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=315#oembed-5

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=315#oembed-6

Altered AI

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=315#oembed-7

Suggested Readings

- Coke's AI Ad
- Game made with AI

Suggested Assignment

Assignment Description: From AI-Generated Essay to TED Talk

Overview:

In this assignment, you will explore the intersection of AI-generated content and public speaking. You will prompt a Generative AI (GAI) to create a 500-word essay on a topic of your choice, within academic, professional, civic, or personal contexts. After reviewing the AI-generated essay, you will revise it, and deliver the content in the style of a TED Talk, using Descript to edit your video. You should edit out any verbal pauses or disfluencies, and edit eye contact. This exercise will help you evaluate the utility of AI in creating engaging and meaningful discourse, and enhance your video editing skills.

Objectives:

1. Utilize AI technology to generate written content on a specific topic.
2. Enhance your editing and content refinement skills by adapting AI-generated text into a compelling spoken presentation.
3. Critically assess the effectiveness of AI-generated content in making insightful points and engaging an audience.
4. Develop skills in video editing and production using Descript.

Instructions:

1. **Select a Topic:** Choose a topic that interests you within the suggested contexts (academic, professional, civic, or personal). For example, “Urban Planning in Houston” as a civic topic or “Trends in Automotive Manufacturing” as a professional topic.
2. **Generate the Essay:** Use a GenAI tool to produce a 500-word essay on your selected topic.
3. **Revise and Transform:** Revise the AI-generated essay to suit a TED Talk-style presentation. Focus on making the language engaging and ensuring that the content makes insightful points without being repetitive.
4. **Record and Edit Video:** Record a video of yourself delivering the revised essay.

5. **AI Video/Audio Effects:** Use Descript and/or AlteredAI to enhance your video. This can include editing out pauses and filler in your speech, and adding AI generated/enhanced voices, and/or stock images and video from Descript.
6. **Reflection:** Write a brief reflection on the process, evaluating how well-suited the GenAI's output was for your purposes, how you adapted the content for your presentation, and your experience using Descript for video editing.

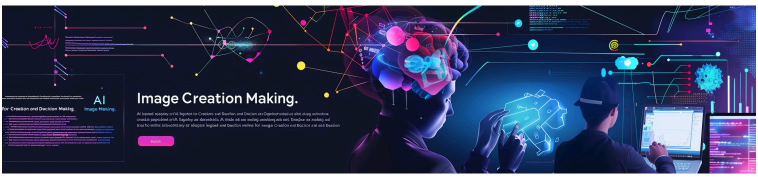
Submission Requirements:

- A copy of the original AI-generated essay.
- The revised script for the TED Talk.
- A video recording of your presentation, edited using Descript (3-5 minutes).
- A written reflection (200-300 words).

Rubric for AI-Generated Essay to TED Talk Assignment

Criteria	Excellent (90-100%)	Good (80-89%)	Satisfactory (70-79%)	Needs Improv (<70%)
Quality of AI Essay	Essay is well-developed, insightful, and closely aligns with the chosen topic.	Essay adequately addresses the topic with a good level of insight.	Essay covers the topic but lacks depth or insight.	Essay do effective address topic or coheren
Adaptation for TED Talk	Presentation is highly engaging, effectively revised, and eloquently delivered.	Presentation is clear and revised well but could be more engaging.	Presentation meets basic standards of clarity and engagement.	Present poorly r or lacks and engager
Video Presentation	Video is professionally edited using Descript, with excellent verbal and non-verbal communication.	Video is clear and well-edited with good communication skills.	Video is adequate but lacks polish in editing or communication.	Video is poorly e with significa issues ir delivery
Reflection and Analysis	Reflection is insightful, providing a deep analysis of the AI's effectiveness, the adaptation process, and the use of Descript.	Reflection provides a good analysis with relevant observations.	Reflection is satisfactory, covering basic thoughts on the process.	Reflecti lacks de showing underst or thou

AI FOR IMAGE CREATION AND DECISION MAKING



[Unedited image created by AI. Errors maintained for demonstration purposes.]

Overview

This section addresses the use of AI for image generation and decision-making processes, highlighting the transformative potential of these tools. The discussion begins with practical tips for generating effective AI images, including being specific, using text-based AI for prompt creation, referencing specific styles or artists, and iterating on prompts. It also touches on ethical concerns, particularly the risks of deep fakes and the labor implications of using AI-generated content, emphasizing the importance of transparency and disclosure when utilizing AI, both to address ethical concerns and to provide clarity for audiences. Additionally, it explores the growing role of AI in strategic decision-making and data analysis, noting that while much of this work currently occurs behind the scenes with proprietary tools, consumer-facing AI tools are rapidly advancing and becoming more accessible. Tools covered include MidJourney and Photoshop.

Videos

AI for Image Creation and Decision Making

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=329#oembed-1

Image Generation

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=329#oembed-2

Photoshop

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=329#oembed-3

Adobe Express

One or more interactive elements has been excluded from this version of the text. You

can view them online here:
https://rotel.pressbooks.pub/datarenaissance/?p=329#oembed-4

Suggested Readings

- *Midjourney Guide, pts 1-4*
- MidJourney Reference Codes
- The AI Images That Shook the Photography World in 2023
- How AI Can Help Leaders Make Better Decisions Under Pressure

Suggested Assignment

Assignment Description: AI-Driven Campaign for Social Change

Overview:

In this assignment, you will analyze a set of posts provided on

Google Drive to understand public sentiments about a social change topic or a product debate. Utilizing AI tools, you will then craft a strategic advocacy campaign that uses AI-generated images to educate, engage, and mobilize the public around this issue or product.

Objectives:

- Develop a deep understanding of public sentiment on a social issue using AI analysis.
- Create a strategic advocacy plan based on sentiment analysis to address the issue effectively.
- Utilize AI tools to generate impactful and persuasive imagery that complements your advocacy strategy.
- Craft a detailed campaign strategy document that effectively communicates your goals and methods to a specified audience.

Instructions:

1. **Analyze Sentiments:**
    - Access the provided posts on Google Drive related to either a social issue of your choosing OR a product debate
    - Use a GAI tool to evaluate the sentiments and opinions on the issue based on uploading at least

25 posts.
- Summarize the findings to guide your campaign strategy.

2. **Select a Topic and Develop Advocacy Strategy:**
    - Choose a specific aspect of the social issue based on the sentiment analysis that you want to advocate for or against.
    - Using GAI, Develop a clear advocacy plan, detailing how you intend to shift or reinforce public opinion using strategic communication.

3. **Generate Campaign Images:**
    - Use an AI imaging tool to create at least four visuals that strongly convey your campaign's message aligned with the findings from the sentiment analysis.
    - Ensure each image is cohesive with the campaign's theme and aesthetically compelling.

4. **Develop a Campaign Strategy Document:**
    - Outline the campaign's goals and objectives based on your advocacy strategy.
    - Define your target audience and explain how your

visuals and messages cater to this group, informed by the sentiment analysis.
    - Describe the anticipated impact of your campaign and how you plan to measure this impact.

Submission Requirements:

- A set of at least 4 AI-generated images.
- A comprehensive campaign strategy document (approximately 500 words).
- A summary and evaluation of the sentiment analysis findings (approximately 300 words).

Rubric for AI-Driven Campaign for Social Change Assignment

Criteria	Excellent (90-100%)	Good (80-89%)	Satisfactory (70-79%)	Needs Improvement (<70%)
Quality and Impact of Images	Images are highly creative, aesthetically compelling, and strongly convey the campaign's message based on sentiment analysis.	Images are well-designed and communicate the campaign's message effectively.	Images are adequate but lack creativity or clear messaging.	Images are poorly designed or fail to communicate the campaign's message.
Cohesion and Theme	All images are cohesive and form a unified visual theme that enhances the campaign's objectives.	Images maintain a general theme with minor inconsistencies.	Some images feel disconnected from the campaign's overall theme.	Images show little to no thematic connection.
Campaign Strategy Document	Document is detailed, well-organized, and includes a thorough analysis of goals, audience, and impact based on sentiment analysis.	Document covers all necessary aspects but lacks depth in analysis.	Document is somewhat vague and missing depth in key areas.	Document is incomplete or poorly organized.

Criteria	Excellent (90-100%)	Good (80-89%)	Satisfactory (70-79%)	Needs Improvement (<70%)
Innovation and Creativity	Shows exceptional creativity and innovation in campaign design and use of AI for images, informed by sentiment analysis.	Demonstrates good creativity and makes a solid effort in utilizing AI tools.	Shows basic creativity and limited use of AI capabilities.	Lacks creativity and does not effectively utilize AI tools.
Use of Sentiment Analysis	Effectively utilizes sentiment analysis to inform and enhance strategic decisions in campaign planning.	Adequately uses sentiment analysis with some impact on strategic decisions.	Uses sentiment analysis but with limited effectiveness in strategy formulation.	Fails to effectively incorporate sentiment analysis into campaign strategy.

AI TOOLS AND APPLICATIONS FOR SOCIAL MEDIA

[Unedited image created by AI. Errors maintained for demonstration purposes.]

Overview

This section explores the integration of AI with social media, emphasizing its impact on businesses across various industries. The discussion highlights the widespread adoption of AI, with a survey from January 2024 revealing that nearly 85% of marketers were using AI in some capacity. It highlights practical applications of GenAI in social media, such as conducting customer research, creating content, enhancing personalization, analyzing data, improving customer service, and optimizing advertising strategies. It explores the AI-based tools, Feedhive, Capcut, and Udio.

Videos

AI Tools and Applications for Social Media

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=331#oembed-1

Feedhive

One or more interactive elements has been excluded from this version of the text. You can view them online here:

https://rotel.pressbooks.pub/datarenaissance/?p=331#oembed-2

Capcut

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=331#oembed-3

Udio

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=331#oembed-4

Song 1:

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=331#audio-331-1

Song 2:

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=331#audio-331-2

Suggested Readings

- *How to Use AI For a More Effective Social Media Strategy*
  (including embedded podcast)
- *ChatGPT-4o Features*

- Generative AI Product Tracker

Suggested Assignment #1:

Assignment Description: Social Media Optimization with Feedhive and AI Analysis

Overview:

In this assignment, you will create a social media post using Feedhive, analyze its predicted performance using AI tools, make adjustments based on the AI's feedback, and then write a reflection on your process and findings. This exercise will help you understand how to use digital tools to optimize social media content strategically. NOTE: You may want to think ahead and use this assignment as part of your larger final project in the last week of the class. It's ok to double-dip.

Objectives:

1. Learn to use Feedhive for social media content creation.
2. Apply AI tools to analyze and predict the performance of social media posts.
3. Enhance social media posts based on analytical insights.

4. Reflect on the effectiveness of AI tools in improving social media content strategy.

Instructions:

1. **Create an Initial Post:**
    - Use Feedhive to design and prepare a social media post relevant to a topic of your choice.
    - Ensure the post is visually appealing and has engaging content that is likely to resonate with your target audience.
2. **Analyze Post with AI:**
    - Utilize Feedhive's AI tool to predict the performance of your post.
    - Take note of any recommendations or insights provided by the AI regarding the optimization of your post.
3. **Make Adjustments:**
    - Revise your post based on the AI's feedback. Consider changes in wording, hashtags, visuals, or timing of the post.
    - Document the changes you make and your

rationale behind each decision.

4. **Write a Reflection:**
    - Reflect on the entire process in a brief essay. Discuss the role of AI in social media strategy, the effectiveness of your adjustments, and any insights you gained about content optimization.

Submission Requirements:

- Screenshots or links to both the original and revised social media posts.
- A reflection essay (250-500 words).

Rubric for Social Media Optimization Assignment

Criteria	Excellent (90-100%)	Good (80-89%)	Satisfactory (70-79%)	Needs Improvement (<70%)
Quality of Initial Post	Post is exceptionally well-crafted, visually appealing, and content is highly engaging.	Post is well-designed and engaging but could have better visual or content elements.	Post is adequate, with decent content and visuals.	Post lacks appeal, with poor content or visuals.
AI Analysis and Adjustment	Adjustments are highly effective, showing a deep understanding of AI feedback and significantly enhancing the post's potential.	Good adjustments, with thoughtful consideration of AI feedback.	Some adjustments made, but not all AI feedback was utilized effectively.	Minimal or ineffective adjustments made, ignoring key AI feedback.
Creativity and Innovation	Shows exceptional creativity and innovation in both original and revised posts.	Demonstrates good creativity and makes a solid effort to innovate.	Shows basic creativity and limited innovation in revisions.	Lacks creativity or significant innovation in post creation.

Reflection and Analysis	Reflection is insightful, providing a deep analysis of the AI's effectiveness and personal learning.	Reflection provides a good analysis with relevant observations.	Reflection is satisfactory, covering basic thoughts on the process.	Reflection lacks depth, showing little understanding or thought.

Suggested Assignment #2

Exploring AI Tools in Communication and Collaboration

Overview:

In this project, you will explore a new AI tool not covered in this class that enhances communication and collaboration, such as a sentiment analysis tool, a chatbot, or a content generation tool. You will create a 10-15 minute video presentation to discuss your experience with the tool, its functionality, potential applications in your industry, and demonstrate the tool in action. Additionally, you will submit a product created using the tool.

Objectives:

1. Gain practical experience with an AI tool that facilitates communication and collaboration.
2. Understand and articulate the tool's functionality and its relevance to specific industries.
3. Explore and demonstrate the tool's potential applications in real-world scenarios.
4. Develop presentation skills in explaining and demonstrating technology.

Instructions:

1. **Select an AI Tool:**
    - Choose an AI tool that interests you and is relevant to your field of study or a potential career path. Ensure the tool can be demonstrated in your presentation.
2. **Experiment and Create:**
    - Use the tool to create a product or output that showcases its capabilities. This could be a report from a sentiment analysis, a conversation with a chatbot, or content generated by an AI.

3. **Video Presentation:**
    - Prepare a 10-15 minute video that includes:
        - An introduction to the tool and why you chose it.
        - A discussion on the tool's core functionalities.
        - An exploration of its potential applications within your industry.
        - A live demonstration of how the tool works.
        - Your personal insights and experiences while using the tool.
4. **Submission Requirements:**
    - A 10-15 minute video presentation
    - The product created with the AI tool (e.g., documents, chat logs, videos).

Rubric for Exploring AI Tools in Communication and Collaboration

Criteria	Excellent (90-100%)	Good (80-89%)	Satisfactory (70-79%)	N In (<
Understanding of Tool	Demonstrates a deep understanding of the tool's functionality and potential applications.	Shows a solid understanding with minor inaccuracies.	Basic understanding shown, with some key functionalities not fully covered.	La un of fu
Demonstration of Tool	Provides an effective, clear demonstration of the tool in action, showcasing its capabilities.	Good demonstration but may lack clarity or depth in showcasing capabilities.	Adequate demonstration but misses opportunities to showcase potential.	Po in de of
Relevance to Industry	Excellently articulates the tool's relevance and potential impact on the industry with insightful analysis.	Effectively discusses relevance to industry with some insightful points.	Generally discusses relevance, but with limited insight or depth.	Fa ef co to in re
Quality of Presentation	Presentation is engaging, well-organized, and professionally executed.	Presentation is clear and structured, with minor issues in delivery.	Presentation meets basic requirements but lacks engagement or polish.	Pr di un un ex

Criteria	Excellent (90-100%)	Good (80-89%)	Satisfactory (70-79%)	N In (<
Innovation and Insight	Shows exceptional creativity and insight in the use of the tool and in drawing conclusions.	Demonstrates good creativity and insight in the application of the tool.	Shows basic creativity; insights are somewhat predictable.	La ar of m in
Product Quality	The submitted product excellently demonstrates the tool's capabilities and aligns with the project objectives.	Product is well-made and demonstrates the tool's capabilities adequately.	Product meets basic requirements but lacks refinement or full capability demonstration.	Pr no de to ca po

LEGAL AND SOCIAL IMPLICATIONS OF AI

[Unedited image created by AI. Errors maintained for demonstration purposes.]

Overview

This section covers recent lawsuits against OpenAI, notably by the Author's Guild and the New York Times, alleging that ChatGPT's training on copyrighted works constitutes copyright infringement, with specific concerns about verbatim reproduction and income loss for authors. The New York Times lawsuit presents a stronger case by demonstrating detailed examples of regurgitated content and arguing that AI-generated summaries impact their revenue. OpenAI defends its practices as fair use, introducing opt-out measures for sites and emphasizing efforts to prevent regurgitation. It covers the GenAI tool RunwayML.

Videos

The Legality of LLM Training

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=335#oembed-1

RunwayML

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=335#oembed-2

Suggested Readings

- Ch. 4 of *Resisting AI: An Anti-Fascist Approach to Artificial Intelligence*
    - If you're interested, the full book is available online through our library, here.
- *OpenAI Creates Realistic Videos*
- *The People Onscreen Are Fake. The Disinformation Is Real.*

Suggested Assignment

AI Ethics Video Project

Overview:

In this assignment, you will create a video using RunwayML and any other AI tools you find appropriate to explore and explain a specific ethics issue in artificial intelligence. This project aims to deepen your understanding of AI ethics and develop your ability to communicate complex ideas effectively through digital media.

Objectives:

1. Investigate and understand a significant ethical issue related to artificial intelligence.
2. Utilize RunwayML and other AI tools to create a compelling and informative video presentation.
3. Enhance digital storytelling skills with a focus on ethical implications in technology.

Instructions:

1. **Select an AI Ethics Issue:** Choose an ethics issue in AI, such as data privacy, algorithmic bias, surveillance, or AI in warfare. Research the topic thoroughly to understand different perspectives and concerns.
2. **Script and Storyboard:** Develop a script that clearly explains the chosen issue, its implications, and potential solutions. Create a storyboard to plan your video's visual and textual content.
3. **Video Creation:**

    - Use **RunwayML** to generate visual elements and effects that enhance the narrative of your video.
    - Incorporate additional AI tools as needed for text-to-speech, background music, or data visualization.

4. **Edit and Finalize:** Compile and edit your video to

ensure it is clear, engaging, and informative. Check that the video effectively communicates the ethical issue and your research findings.

Submission Requirements:

- A video of 3-5 minutes explaining the selected AI ethics issue.
- A brief document (300-500 words) outlining your script, storyboard, and a description of how you used AI tools in the project.

Rubric for AI Ethics Video Project

Criteria	**Excellent (90-100%)**	**Good (80-89%)**	**Satisfactory (70-79%)**	**Needs Improv (<70%)**
Content Accuracy and Depth	Provides a thorough and accurate explanation of the AI ethics issue, with detailed examples and well-researched perspectives.	Adequately explains the AI ethics issue with some good examples and research.	Provides a basic explanation of the issue, but lacks depth or detail.	Explanat unclear c incorrect lacks rele or accura
Use of AI Tools	Creative and effective use of RunwayML and other AI tools to enhance the video's impact and clarity.	Good use of AI tools that somewhat enhance the video's educational value.	Basic use of AI tools; enhancements are minimal or only slightly effective.	Limited ineffecti of AI too does not contribu video qu
Engagement and Presentation	Video is highly engaging, visually appealing, and excellently presented; maintains viewer interest throughout.	Video is engaging and well-presented with minor areas for improvement.	Video is somewhat engaging but could be more dynamic or visually appealing.	Video is poorly presente engaging difficult follow.
Technical Quality	Video is technically impeccable with professional-quality editing, sound, and visuals.	Video has good technical quality with some minor errors in editing or sound.	Video meets basic technical standards but shows areas needing improvement.	Technica issues significan detract f the video overall q

AI ETHICS: PRIVACY, SECURITY, AND BIAS

[Unedited image created by AI. Errors maintained for demonstration purposes.]

Overview

GenAI faces significant ethical challenges, including biases in training data that can lead to discriminatory outputs and issues with content moderation. AI is increasingly being used in critical areas like healthcare and decision-making, which can result in biased or harmful outcomes. Additionally, AI's environmental impact is substantial, contributing to high levels of carbon emissions and water consumption. Tools covered include Zotero, Consensus, and Elicit.

Videos

AI Ethics

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=339#oembed-1

Zotero

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=339#oembed-2

Consensus and Elicit

One or more interactive elements has been excluded from this version of the text. You can view them online here: https://rotel.pressbooks.pub/datarenaissance/?p=339#oembed-3

Suggested Readings

- Covert Racism in LLMs (Overview, full paper linked)
- John Oliver on AI
- $2 Workers that Made AI Safe

Suggested Assignment

AI Ethics Literature Review on AI in Academic Writing

Overview:

In this assignment, you will explore the ethics of using artificial intelligence (AI) tools in academic writing. Your literature review will specifically address the question: "Is it ethical to use AI for academic writing?" You must also incorporate Fitchburg State University's academic integrity policy into your analysis to contextualize the ethical considerations within your institutional framework.

Objectives:

1. Critically examine and synthesize existing research on the use of AI in academic writing.
2. Evaluate the ethical implications of using AI tools for academic purposes.
3. Understand and apply Fitchburg State University's academic integrity policy to your analysis.
4. Develop a well-supported argument based on your review of the literature.

Instructions:

1. **Literature Search and Analysis:**

    - Use **Elicit** to gather scholarly articles and papers on AI ethics, with a focus on its use in academic writing.
    - Apply **Consensus** to synthesize findings and discern major themes and ethical considerations.

2. **Writing the Literature Review:**

    - Your review should specifically address the ethical implications of using AI in academic settings.
    - Incorporate a discussion on Fitchburg State University's academic integrity policy, analyzing how it relates to the use of AI tools.
    - Organize the literature into themes or categories that support your analysis and conclusion regarding the ethics of AI in academic writing.

3. **Citation and Referencing:**

    - Cite all sources accurately using a citation style approved by your department.

4. **Questions to Consider (from *AI and Writing*):**

    - Some claim that every GenAI output is inherently plagiarism, on grounds that everything the GenAI creates was originally someone else's words or thoughts. Yet, similarly, most of what we say, write, and think could be aruged to be plagiarism, because we so often cull it, reorder it, and re-present it from other peoples' ideas and expressions. First, consider the ramifications of this on how we think about things like individuality, sense of self, expression, creativity, and even intelligence. Second, consider how such an acknowledgment of the similarities between GenAI and human thinking might affect what we understand about plagiarism and academic integrity.

- When you write, it's likely that you already use some AI tools such as spell checkers and grammar checkers. These tools have altered our learning in many ways. You no longer need to know how to spell every word correctly. You no longer *need* to know all of the rules of grammar. Is using these kinds of tools a violation of academic integrity? Write a short essay comparing the accepted use of AI tools such as spelling and grammar checkers with the often-prohibited use of GenAI.
- Is monitoring for plagiarism an act of policing or an act of education?

Submission Requirements:

- A literature review document of 1500-2000 words addressing the specified topic.
- A bibliography listing all referenced sources.

Rubric for AI Ethics Literature Review on AI in Academic Writing

Criteria	Excellent (90-100%)	Good (80-89%)	Satisfactory (70-79%)	Needs Improvem (<70%)
Relevance and Depth of Research	Exceptionally relevant sources are used; provides deep insight into ethical considerations of AI in academic writing.	Sources are relevant and provide good coverage of the topic with substantial insight.	Sources cover the topic adequately but lack depth or broader insight.	Sources are insufficient relevant or to adequat cover the topic.
Integration of FSU Academic Integrity Policy	Excellently integrates FSU's academic integrity policy, providing a strong ethical framework for the analysis.	Adequately integrates the academic integrity policy with good relevance to AI ethics.	Mentions FSU's policy but with limited integration or relevance to AI ethics.	Fails to effectively incorporat relate FSU' academic integrity policy.
Use of AI Tools	Effectively uses AI tools to enhance the breadth and depth of literature analysis.	Uses AI tools well, contributing positively to the analysis.	Uses AI tools adequately, but integration into research is basic.	Uses AI to minimally ineffectivel does not enhance research.

Criteria	Excellent (90-100%)	Good (80-89%)	Satisfactory (70-79%)	Needs Improvem (<70%)
Organization and Clarity	Review is exceptionally well-organized and clearly presents a coherent narrative regarding the ethics of AI in academic writing.	Well-organized and presents clear findings with minor issues.	Somewhat organized but lacks clarity in presenting findings.	Poorly organized a difficult to follow.
Critical Analysis and Argumentation	Provides a deep, insightful analysis with a strong, well-supported argument regarding the ethical use of AI.	Solid analysis and argumentation with some insightful observations.	Basic analysis with an adequate argument but lacks depth.	Lacks critic analysis or to form a coherent argument.
Citation and Referencing	All sources are cited flawlessly according to the recommended citation style.	Minor errors in citation style but generally well done.	Citation style is inconsistent or has several errors.	Many citat errors or incorrect u of citation style.

THE FUTURE OF AI AND JOBS

[Unedited image created by AI. Errors maintained for demonstration purposes.]

Overview

We've covered a lot of ground in this chapter, from practical use of AI tools to understanding their societal, legal, and ethical implications. As we wrap up, I have two tasks for you: first, find a new AI tool not covered in this class, explore its features, and evaluate its potential use in your workflow. Second, update your resume to reflect your new AI skills and use AI to help tailor it, along with a cover letter, for a job you find interesting. Thank you for joining me on this journey; it's been a pleasure discussing these critical issues and tools with you.

Wrap-up

Suggested Readings:

- Using AI for resumes
- Smart Glasses Tell You What to say on first date

Suggested Assignment #1

AI-Enhanced Resume and Cover Letter Creation

Overview:

In this assignment, you will utilize AI tools to update your resume and craft a cover letter for a potential job that requires AI skills. The goal is to effectively integrate AI technology to enhance the presentation of your skills, experience, and educational background, and to articulate how these align with the job requirements.

Objectives:

1. Demonstrate the ability to use AI tools to create

professional and polished job application materials.

2. Highlight AI skills and experiences in a way that is tailored to the job description.
3. Develop a cover letter that effectively complements the resume and persuasively communicates your qualifications.

Instructions:

1. **Select a Job Posting:**
   - Find a job posting that interests you and requires AI skills. This job should be relevant to your career aspirations.
2. **Update Your Resume:**
   - Use an AI-powered tool (like Canva's Resume Builder or another reputable AI resume tool (you can use ChatGPT or Claude) to update your resume. Ensure it is visually appealing and organizes your information in a way that highlights your most relevant experiences and skills.
   - Specifically focus on detailing any AI-related coursework, projects, or work experiences.
3. **Craft Your Cover Letter:**

   - Utilize an AI tool designed to assist in writing, such as Grammarly or an AI writing assistant, to create a cover letter that addresses the specific job description.
   - Your cover letter should introduce who you are, highlight your AI skills and experiences, explain why you are a good fit for the position, and show your knowledge of the company.

4. **Submission:**

   - Submit both the updated resume and the cover letter as PDF files.

Submission Requirements:

- A job posting that requires AI skills.
- An updated resume tailored to the job posting.
- A cover letter addressing the specific qualifications and requirements of the job.

Rubric for AI-Enhanced Resume and Cover Letter Creation

Criteria	Excellent (90-100%)	Good (80-89%)	Satisfactory (70-79%)	Needs Improver (<70%)
Relevance and Tailoring	Resume and cover letter are exceptionally tailored to the job, highlighting relevant AI skills and experiences.	Documents are well-tailored with relevant details but could be more specific.	Documents meet the basic relevance but lack customization.	Documen not adequ match the requireme highlight relevant sk
Use of AI Tools	Uses AI tools effectively to produce polished and professional documents.	Uses AI tools well, with minor errors in document formatting or language.	Adequate use of AI tools, with noticeable formatting or linguistic issues.	Poor or incorrect u AI tools; document profession
Presentation and Formatting	Documents are visually appealing, well-organized, and error-free.	Documents are generally well-presented but could use slight improvements.	Documents are adequately presented but lack careful formatting.	Documen poorly formatted difficult to or contain errors.
Persuasiveness and Content	Cover letter is compelling, clearly articulating the candidate's strengths and fit for the role.	Cover letter is persuasive but lacks impactful language or specific examples.	Cover letter conveys basic suitability but is not compelling.	Cover lett fails to per or properl address th requireme

Criteria	Excellent (90-100%)	Good (80-89%)	Satisfactory (70-79%)	Needs Improver (<70%)
Creativity and Insight	Shows creativity in presenting information and insightful understanding of job requirements.	Demonstrates good effort in creativity and understanding, with minor lapses.	Shows basic creativity; understanding of job requirements is adequate.	Lacks crea and does r demonstra understan of job requireme

Suggested Assignment #2

Hypothetical Social Media Campaign Planning

Overview:

For this project, you will develop a detailed plan for a one-week advertising campaign for a product of your choice. This plan will be hypothetical and won't involve actual posting. You will utilize Generative AI (GAI) tools to help create a comprehensive document that outlines your campaign's audience, keywords, suggested platforms, and creative ideas for posts.

Objectives:

1. Use GAI tools to aid in the strategic planning of a social media campaign and creation of content.
2. Develop a deep understanding of target audience analysis and keyword optimization.
3. Plan content types and choose appropriate social media platforms for campaign deployment.
4. Enhance planning and creative thinking skills in the context of digital marketing.

Instructions:

1. **Campaign Concept and Planning:**
    - **Select a Product:** Choose a product or service to advertise. This could be an actual product, a hypothetical product, or this class.
    - **Define Campaign Goals:** Set clear, measurable objectives for what you aim to achieve with your campaign.
2. **Use of Generative AI Tools:**
    - **Audience Analysis:** Use a GAI tool to help define your target audience. Identify demographics, interests, and behaviors that will influence your

campaign strategy.

- **Keyword and SEO Optimization:** Utilize GAI to generate keywords that will help optimize your content for search engines and social media search algorithms.
- **Platform Selection:** Determine the best social media platforms for your campaign based on the audience and content type. Use GAI to gather data on platform demographics and effectiveness.

3. **Content Strategy Development:**

    - **Content Ideas:** Generate at least one photo post and one video post using GAI tools. Consider how these ideas can engage your defined audience.
    - **Scheduling Strategy:** Plan the timing of your posts. You can use GAI to suggest optimal times for posting based on user activity and engagement data.

4. **Campaign Document Creation:**

    - **Compile a Campaign Strategy Document:** This document should include your audience analysis, keyword list, chosen platforms, and your posts. Describe how each element of your plan aligns with your overall campaign goals.

Submission Requirements:

- A comprehensive campaign strategy document (about 1000 words) detailing your audience, keywords, platform selection, and content ideas.

Rubric for Hypothetical Social Media Campaign Project

Criteria	Excellent (90-100%)	Good (80-89%)	Satisfactory (70-79%)	Needs Improvement (<70%)
Strategic Planning	Demonstrates outstanding strategic thinking with innovative objectives and thorough audience analysis.	Shows solid planning with well-defined objectives and good audience analysis.	Basic planning evident with general objectives and limited audience insight.	Lacks clear objectives or understanding of the target audience.
Use of GAI Tools	Uses GAI tools creatively and effectively to enhance all aspects of the campaign planning.	Uses GAI tools appropriately with good effect on the planning process.	Adequate use of GAI tools, but integration lacks creativity or impact.	Poor or incorrect use of GAI tools, with little to no benefit to planning.
Content Strategy and Innovation	Content ideas are exceptionally engaging, well-designed, and perfectly tailored to the audience.	Content ideas are engaging and suit the audience, with minor areas for improvement.	Content ideas meet basic standards but lack creativity or polish.	Content ideas are ineffective, poorly designed, or do not engage the audience.
Quality of Campaign Document	Document is comprehensive, well-organized, and includes detailed analysis and planning.	Document covers all necessary aspects but could be more detailed or organized.	Document is adequate but lacks detail and organization in key areas.	Document is incomplete, poorly organized, or lacks critical information.

Header Image by J.J. Sylvia IV using MidJourney is licensed

WRAP-UP

Key Takeaways

- The ethics of AI matter greatly and can have significant consequences, as demonstrated by the controversy surrounding Sports Illustrated's use of AI-generated content without proper disclosure, which damaged the brand's reputation.
- AI has the potential to dramatically increase productivity and output quality across various fields, from writing and coding to image and video generation, but it is essential to carefully review and edit AI-generated content before using it.
- Prompt engineering is a crucial skill in getting the best results from AI, as the way you craft your prompts can significantly influence the output; providing clear instructions, context,

and iterative feedback can help optimize AI performance.

- The rapid advancement of AI raises important philosophical questions about the nature of art and the role of human creativity, as well as concerns about job displacement, cognitive deskilling, and the potential for AI to be weaponized for surveillance and warfare.
- To harness the benefits of AI while mitigating its risks, it is vital to proactively consider and address these ethical concerns through the development of safety mechanisms, laws, and international agreements that prioritize transparency, accountability, and the protection of human rights.

Exercises

1. Reflect on a time when you encountered AI-generated content online. Did you initially

realize it was AI-generated? How did this realization affect your perception of the content and the platform? Discuss the importance of transparency in AI content creation.

2. Choose a specific AI tool or application mentioned in the chapter (e.g., AI-assisted writing, image generation, video editing) and consider how it could be integrated into your personal or professional life. What benefits and challenges do you anticipate? How would you ensure the ethical use of this tool?
3. Imagine you are tasked with creating a set of ethical guidelines for a company developing AI technologies. What key principles would you include to ensure the responsible development and deployment of AI? Consider issues such as transparency, accountability, privacy, and fairness.
4. The video mentions the potential impact of AI on jobs and the workforce. Select an industry or profession and analyze how AI might disrupt or transform it in the coming years. What skills do you think will be most valuable in this new landscape, and how can

individuals and organizations prepare for these changes?

PART XII
SOPHIA DISCUSSION GUIDES

Dr. Sylvia's Communication Law & Ethics courses have successfully partnered with SOPHIA (Society of Philosophers in America) and the Douglas and Isabelle Crocker Center for Civic Engagement to create robust discussion guides aimed at fostering meaningful dialogue in the community. SOPHIA's mission revolves around employing philosophical inquiry to improve people's lives and build community. To this end, they aim to be inclusive, avoiding jargon and offering simplified, quick explanations for the ease of public engagement. Students in this course have tailored their conversations to issues related to ethics and media and held several off-campus public discussions each time the course is offered.

By aligning with SOPHIA's mission and leveraging the resources and community reach of the Crocker Center, Dr. Sylvia's courses have created an enriched, accessible platform for community dialogue on communication law and ethics. Several of the discussion guides are included below, for potential use either in class, or as part of your own public discussion.

ETHICS OF SEARCH ENGINES

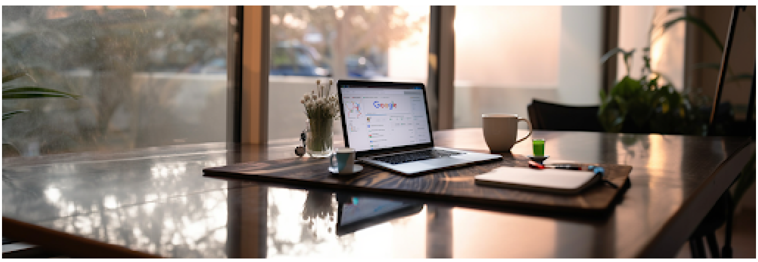


Figure 1: Laptop Search, generated by MidJourney.

In today‘s digital world, we use search engines everyday to find information we are looking for. Whether that be for research, news, entertainment, shopping or any general curiosity we may have, we trust that these engines will provide us with the best results. However, how much should we trust these results?

To understand why we get the results that we do, we have to look at how major search engines like Google, Yahoo, and Bing engineer their algorithms. In order for a site to come up on the results page, it must be linked to certain keywords or other related sites. The more links and keywords associated with the site the higher on the results page it will show up. Companies can also pay to have their sites show up at the top of results. These are labeled as "ads" and are separated from the organic search results. Search engines also rely on algorithms that assess users' data to personalize search results.

Major search engines have been criticized for giving inappropriate, racist, and untruthful results. Some argue these engines are ethically responsible for the information they provide, given their high influence on the everyday person. Others may say it is the user's job to be able to distinguish and judge the information provided.

There are many layers to this issue as we will discuss. Ultimately we look to answer these key questions: Are major search engines' current filters and algorithms ethical? Should they be doing more? Or less? And for what ethical reasons?

I . Key Questions

1. When one searches for something through a search engine, they expect to receive all the related information that pertains to their inquiry. Are there reasons it may be more ethically responsible to withhold some information? For instance, can you think of some reasons search engines might withhold false or inaccurate content? What should be the deciding factor when determining what should be restricted in search results?

2. What should be the standards for filtering search results and who should set them? Should it be the government's responsibility to set policies for search engine filters, or should it be the individual search engine companies that decide on what content is shown when you search on their website?
3. Search engine users have grown to trust their search engines even when, perhaps, they shouldn't. For example, the founders of Google, Larry Page and Sergey Brin wrote a paper proposing the core ideas for the search engine while they were students at Stanford. They argued that advertising would inherently corrupt a search engine, biasing it toward advertisers and away from the needs of consumers (Foroohar, 2019). What are the ethical impacts of ads in search results?
4. Should search engines set standards to censor content for children, similar to television? Why or why not?

II. Mini Prompts

1. In 2016, the Washington Post published an article about Google receiving backlash for its image search results. At the time, if you searched "three white teenagers" on Google Images, you would mostly see stock photos of white teenagers. But if you searched "three black teenagers", you would get multiple mugshots of young

African Americans. How should search engines address this issue, especially in light of increased calls to address systemic racism?

2. The algorithmic practice of personalization in which a search engine looks at users' location and previous searches, is a great way to get search results that pertain to what one is particularly interested in. However, this can lead to some pitfalls. For example, if Google is aware of your political alignment, ideologies, or any other opinion you may hold through your searches, they could show you results that fit your existing perspectives. Many fail to see the other side of issues due to the one-sided information they are receiving. This phenomenon is often referred to as "autopropaganda." Many argue it is also anti-democratic in nature, as a good democracy "requires citizens to see things from one another's point of view," (Pariser, 2011). Since 2011, Google has begun reducing the extent of their personalization algorithm. How might we make sense of the ethics of personalized search results in terms of filter bubbles and the impact on democracy? What other ways should search engines address this, if at all?
3. The algorithmic practice of Search Engine Optimization (SEO) is a system in which sites can better their chances of being at the top of results by linking their site to keywords and related sites. Search engines have set rules that limit the extent of this optimization by, for example,

penalizing the ranking of sites using keywords that do not pertain to the content of their site. Are these rules productive? In what ways do they protect users? What are some additional ethical considerations related to companies using SEO practices to their advantage?

This one-sheet was created for the SOPHIA of Worcester County chapter by students in the Communication Law and Ethics course at Fitchburg State University and edited by Dr. J.J. Sylvia IV and Dr. Kyle Moody. Its creation was supported by SOPHIA and the Douglas and Isabelle Crocker Center for Civic Engagement. Students included: Alexander Pierre, Maximillian Simonelli, Emma Jacques, Kevin Sim, John Javaloyes, Ryan Titemore, and Colby Molleo. Image generated by MidJourney.

COVID-19: SURVEILLANCE AND PERSONAL PRIVACY

Overview:

The unprecedentedly pervasive and deadly nature of the COVID-19 pandemic has led to an equally unprecedented expansion of public health surveillance. Such expansion is primarily geared toward technological advances, such as tracking apps, as preventative measures in combating the virus' spread. Though seemingly implemented with the best of intentions – i.e. for the benefit of the population at large against a deadly, globe-spanning disease – the increased technological surveillance proposed and/or implemented by various governments have raised legal and ethical concerns over the breach of personal liberties, namely privacy. For many, the governmental overreach inherent in such surveillance undermines individual, inalienable human rights (namely privacy of information) that are fundamental to our democratic society.

Preliminary Discussion Questions:

1. Do you think that the collective good provided by increased technological COVID-19 surveillance supersedes certain privacy rights? Why or why not? What moral obligation do you feel you have toward your community and how far does it extend?
2. Personally invasive security measures have been a regular facet of contemporary society due to increased attention toward public safety. What circumstances would allow for you to comfortably forfeit certain personal liberties, e.g. privacy of information, for the sake of public safety (if you ever would)? If no such situation exists, why not?

Part I: Tracking Apps / Contact Tracing

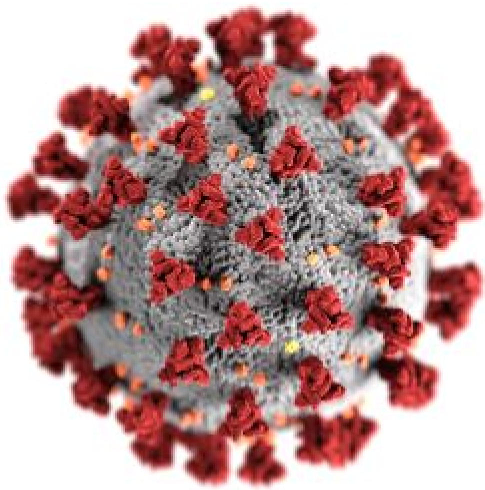


Figure 2: A picture of the COVID-19 disease, as seen under a microscope.

To counter the spread of the COVID-19 virus, several governments and organizations across the globe have proposed and/or introduced new methods of technological surveillance. The World Health Organization has implemented enhanced surveillance tactics through wearable technologies, such as bracelets or watches, which allow public health authorities to gauge people's temperatures and/or other indicators of potential COVID-19 symptoms; The United States and China have enlisted the aid of advanced, QR-code-carrying drones, capable of monitoring the temperature, heart rate, respiratory abnormalities of, and distance between individuals; Russia has implemented artificial-intelligence facial-recognition software, connected to a nationwide network/ camera system, allowing authorities to identify, locate, and apprehend individuals in violation of quarantine/social distancing protocols within a 30-minute window; and Taiwan has created a geofencing system, alerting authorities if cell phone users stray beyond designated quarantine zones, as well as if individuals' devices have been turned off or drained of battery life.

Questions for Part I

1. What are your thoughts on the progressions of these digital tools that have made modern contact tracing possible? Is the safety gained by the implementation of technology worth the ethical loss of personal liberties? Why or why not?
2. While much of the more extreme forms of surveillance listed above (i.e. A.I. facial recognition, geofencing, etc.) have not taken root in the United States, many local and state governments have pushed for more advanced surveillance technologies. Cities in New Jersey and Connecticut have successfully secured drones with automated voice messages to enforce social distancing protocols. How are/aren't the nature of these

technologies characteristic of a democratic republic?
3. What more ethically acceptable alternatives, if any, could governments implement to help track and maintain the spread of COVID-19?

Part II: Discrimination

Ethical questions have arisen regarding a correlation between increased COVID-19 surveillance and discriminatory abuses of said surveillance. According to the *Health and Human Rights Journal*, minority groups are at a particularly high risk of incurring infringements of personal privacy during periods of heightened governmental security. Examples of such are cited as early as the eugenics laws of Nazi Germany during World War II, justified as state-healing public health measures; and as late as homophobic abuse targeted at members of the LGBTQ community in South Korea as a result of contact tracing for COVID-19 outbreaks. Despite this not being the intended outcome, many argue that such mistreatment is characteristic of overzealous technological surveillance.

Questions for Part II:

1. Courts in South Korea are known for their refusal to recognize same-sex partnerships, leading many to believe that COVID-19 protocols have served to radically aid

systemic abuses of the LGBTQ community in the country. For example, as part of contact tracing efforts, the government has released detailed information including age, gender, and workplace. How could increased surveillance be used as a tool to perpetuate systemic social prejudices? How might governments more ethically balance the competing needs of contact tracing and privacy?

2. Existing ethical frameworks for public health interventions require evaluation for policies that restrict individual liberty. Two of those questions for evaluation include, "Are the benefits and risks of the public health intervention equitably distributed?" and "Is the intervention the least restrictive alternative for achieving the public health goal?" (Lo and Sim, 2021). How might we assess contract tracing in light of these ethical standards?

Special Thanks to Strong Style Coffee and the continued support of SOPHIA and the Douglas and Isabelle Crocker Center for Civic Engagement.

Eric Bielakiewicz, Grace Bowen, Ryan Esteves, Jack Harney

Dr. J.J. Sylvia's Communication Law and Ethics Course

Figure 2: A picture of the COVID-19 disease, as seen under a microscope. CC0 1.0 Universal (CC0 1.0) Public Domain Dedication

AI AND ETHICS: A DISCUSSION

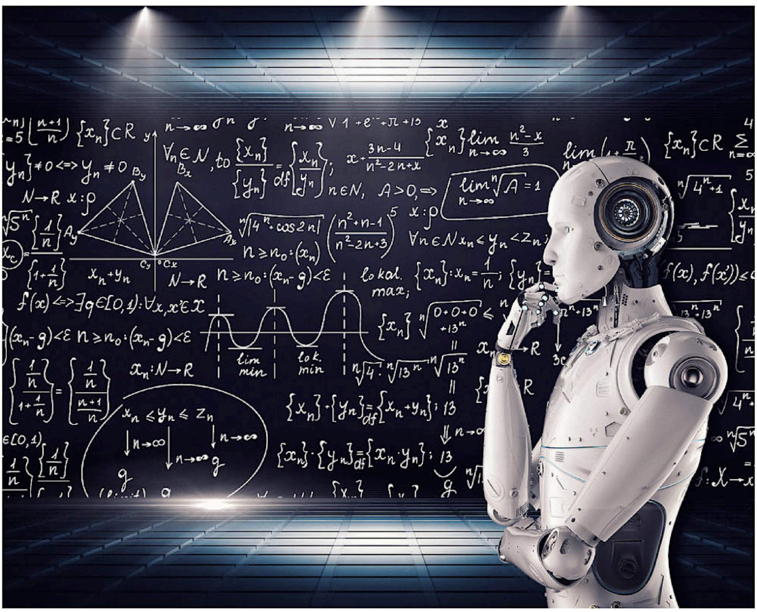


Figure 3: “Artificial Intelligence & AI & Machine Learning”

Artificial intelligence (AI) suggests that machines will one day have the potential to imitate human behavior to complete complex tasks without human assistance. Many modern devices and appliances strive to operate in such a way. AI is not limited to robot technologies or self-driving cars, as it includes software like Siri on the iPhone, home systems, social media, and even many children's toys. With AI integration growing more commonplace, scientists and modern philosophers worry how this might affect consumers.

MIT Professor Sherry Turkle studies the interaction and relationships between humans and devices. In her book Reclaiming Conversation: The Power of Talk in a Digital Age (2015) she examines the comfort people find in simple relationships with their devices and the effects of such. Turkle argues this leads to simpler conversations that are almost transactional in nature. These lusterless conversations lack the means to foster any sort of empathy. Turkle sees the importance of AI, but believes excessive exposure and connection to one's devices are detrimental to human socialization.

Discussion Questions

1. What is the difference between AI software and hardware? How do they operate?
2. What parts of social media platforms use an AI component and what are the dangers of that?
3. Where else do we see AI technologies and software in our everyday lives?

4. Are some AI technologies safer than others? Which, and why?
5. What sorts of protections should be put into place to protect consumers from potential negative aspects of AI systems?
6. In cases of algorithms being made for AI systems, how are fairness and good ethics guaranteed, especially when private corporations are immune from public scrutiny?

AI and Privacy

Operating in a social-digital age, personal information is all the more accessible. Helen Nissenbaum, well recognized for studies in privacy and her concept of "contextual integrity", wishes to create a system that appropriately delegates the use of personal data. With contribution from her collaborators, Nissenbaum has created a series of web plugins including TrackMeNot, Adnostic, and AdNauseam. These are "obfuscating" plugins that interfere with various data collection and ad services.

Question: Why might people worry about private companies having access to their personal data and information? What should private companies be able to do with this private information? What sorts of laws should be proposed, specifically in terms of privacy?

Bias in AI

AI systems can demonstrate bias. Some bias is not actually programmed into the code intentionally, but is the result of user interaction. Helen Nissenbaum uses Google's behavioral advertising system as an example to explain this behavior. If one were to search two different names, one traditionally Caucasian and one traditionally African-American, searching the traditionally African-American name would yield more advertisements for background checks. Because background check advertisements are more likely clicked on when users search traditionally African-American names, Google's system places more ads on searches for African-American names. Thus, racial bias is introduced by the user into the AI system.

Question: What sort of problems could the public face with human bias in AI programming? What kinds of safeguards should be put in place to ensure bias-free AI programming?

Predictive Policing

Only recently surfaced, the New Orleans police department started using a predictive policing program developed by Palantir Technologies in 2012. Palantir had access to personal information including social media data, phone numbers, addresses, licenses, court filings, and more. The software would use these records and private information to predict and deem people potential aggressors or victims. Palantir did this without the consent or knowledge of the City Council.

Question: Does this action by a private company seem like a violation of the law? What are the possible implications of government organizations using AI technology to police its citizens?

This one-sheet was created for the SOPHIA of Worcester County chapter by students in the Communication Law and Ethics course at Fitchburg State University and edited by Dr. J.J. Sylvia IV and Dr. Kyle Moody. It was hosted by Strong Style Coffee and its creation was supported by SOPHIA and the Douglas and Isabelle Crocker Center for Civic Engagement. Students included Miguel Aguiar, Colin Ahearn, Andrew Allen, Ben Bursell, Olivia Grant, Rebecca Landry, Kevin Newey, Martha Melendez, Shane Muir, Edgar Mutebi, Scott Ryan, Ben Sharple.

THE ETHICS OF FAKE NEWS

Figure 4: Fake News Ethics, MidJourney image.

Overview

Former President Donald J. Trump popularized the term "Fake News" during the 2016 U.S. election. Although this term is now used frequently by politicians and the media, much confusion remains over the meaning of the term and what actually "counts" as fake news. The *UNESCO Handbook on Journalism, 'Fake News,' & Disinformation* distinguishes between three categories that include:

- Mis-information: false connection or misleading content
- Dis-information: false context, imposter, manipulated or fabricated content

- Mal-information: Some leaks, harassment, or hate speech

However, some questions still remain about these categories. For example, President Trump sometimes seems to use the term to refer to reporting that he doesn't like. Addressing these definitional questions might help us better understand other ethical issues such as where fake news comes from, who creates it, and whether it is being spread intentionally or accidentally.

Discussion Questions

1. How should we define fake news?
    - What is the difference between real and fake news? For example, how might we draw a line between false statements and political spin? Is there a difference?
    - What benefits might be had from using more specific terms such as mis- or disinformation?
    - Do these terms leave anything out? If so, what?
    - How do you recognize it? Do you know it when you see it?
    - Have you believed something you later found out to be fake news or misinformation?
2. How does fake news spread?

- Who is sharing it? Who is producing it? Why?
- Do you think it's possible to prevent its spread? If so, whose responsibility is it?
- Does social media contribute positively or negatively to the spreading of fake news?

3. Is fake news a *new* problem?

    - Are there historical parallels to the problems we associate with fake news?
    - Is there something unique about fake news that makes it different from these historical parallels?

4. Who *can* we trust?

    - Are there any news sources that you can trust at face value, without additional verification?
    - What strategies do you or should you take before sharing news on social media?

5. How much influence do public figures have in spreading fake news?

    - How effective is the influence of political figures? Professional athletes? Celebrities? Politicians?
    - Is there a correlation between social status and the potential to spread fake news? Do figures take their

reputation into account when making a statement?

Mini-Prompts

> "Because the tools that the public relies on to gauge truth, fairness, and accuracy are designed around the codification of sentiment and the monetization of attention, the 'fake news' battle cannot be won at the level of content alone. 'Indisputable facts play only a partial role in shaping the framing words and images that flow into an audience's consciousness,' notes Entman (2007). Given this scenario, objectivity, while important at the reporting level, is less valuable for establishing trust between news organizations and audiences in the 'fake news' era. As more actors opt to go 'direct' to their audiences using platforms like Twitter, news organizations will be forced to 'follow the conversation' instead of leading the way to establish narratives that accurately inform the public through their reporting. In this regard, as Richard Tofel argues, 'publishing [news] and then fact checking is not enough,' (2015)." – Jonathan Albright, Columbia University

Question: Do you find yourself relying more on the judgment of news outlets or social media? Do you trust one more than the other? What are the advantages/pitfalls of both?

> "Many cannot even tell these days which sources are biased. And if one believes that all media are biased, perhaps it makes less difference to choose an information source that is biased in one's favor. Those who have provided charts

> that attempt to measure the reliability of various media sources since the [2016] election have been met with threats of bodily harm." — Lee McIntyre in Post-Truth

Question: Much of the media literacy efforts in the United States are oriented around helping students build their skepticism toward all information sources, deconstructing it, and asking detailed questions about its source and veracity. How might this pedagogical model of skepticism, built around helping students see the bias in all information sources, impact the way we consume information?

This one-sheet was created for the SOPHIA of Worcester County chapter by students in the Communication Law and Ethics course at Fitchburg State University and edited by Dr. J.J. Sylvia IV and Dr. Kyle Moody. Its creation was supported by SOPHIA and the Douglas and Isabelle Crocker Center for Civic Engagement. Figure 4: Fake News Ethics, MidJourney image, Attribution-NonCommercial 4.0 International

THE ETHICS OF SOCIAL MEDIA USE BY CHILDREN

Figure 5: Kids Using Social Media, MidJourney image

COVID-19 forced us all to turn online for work, school, and social needs. Over the course of 2020 the concepts of screen fatigue and burnout became ubiquitous due to the social restraints of COVID-19, and it is hard to imagine that this is in no way connected with the uptick in the unyielding presence of social media. Smartphones have made the internet portable, and with that we have the ability to connect with others and access all of the information in the world in our back pockets.

This also means that the reach of school and work extends beyond campuses and office buildings as work and school emails can now follow us home. Moreover, there is a bottomless well of content just waiting for us; there are more movies and T.V. shows that could be streamed in a lifetime across more streaming services than you can count on two hands, and that doesn't even take YouTube into account. It is easy to lose yourself scrolling through TikTok, Twitter, or Instagram for hours at a time if you are not mindful of your consumption.

However, social media is not just for consuming content, it also allows us to create content with which others can interact. We are able to broadcast our thoughts in an instant, whether or not we have taken the time to reflect on what we are publishing for the world to see.

The necessity of self-regulating social media intake and what is posted online becomes more acute when discussing the use of social media among children. Use of social media platforms has become an integral part of childhood socialization, and that presents its own unique ethical challenges and questions. Complete abstinence from these platforms means cutting off an entire avenue of social interaction, but unfettered access to social media platforms has its own risks, which begs the question: how might ethics guide us in thinking about children's online presence and access?

Discussion Questions

1. How do you think social media will affect the children of this generation as they grow up? Do the pros outweigh the cons?
2. What do you think about parents who monitor their kids' social media and messages? Is this ethically beneficial for the kids or is it a problematic invasion of privacy? Why?
3. Would you consider social media to be addictive? Why/why not? How much time on social media per week/day would you consider healthy for a kid? How long do you spend on your phone daily?
4. What age or maturity level do you feel is appropriate to start using social media?
5. What are some risks that children may come across when using social media?
6. What are some benefits that children could gain from social media use?

Mini-Prompts

1. Netflix's current "kids only" feature was designed to help kids navigate the vast array of streaming options, some appropriate and many not. Last Spring, Netflix rolled out an even more detailed parental control guide to reinforce

child privacy settings on their platform amidst the exponential increase in screen time due to the COVID-19 lockdown. The "kids only" feature raises the question of if it would be beneficial to implement similar parental controls on other social media apps such as Facebook, Instagram, and TikTok. Moreover, in order for this feature to be rolled out on other social media platforms, developers would need to determine the most ethical way to design its functionality. For example, would a parent or guardian be the one setting up these features? How would transitioning from a "kids only" version of the platform to the full functionality work? Would that transition happen automatically when the user reaches a certain age? Would users of the full website be allowed to interact with users on the "kids only" version? How and why would developers set each of these parameters?

2. In 2000, Congress passed the Children's Online Privacy Protection Act (COPPA). The act "imposes certain requirements on operators of websites or online services directed to children under 13 years of age." Under COPPA, social media services must require all users to be at least 13 years old in order to utilize their platform. Over the last twenty years, not only have these services become more universal, but they have also had an increased presence in daily social interactions. Due to the changes in how social media websites operated in 2000 and the ways they are used today, is 13 still an adequate minimum

age requirement for social media use? Was 13 an adequate age in 2000 when COPPA was first passed?

PART XIII

DATA FEMINISM: THE NUMBERS DON'T SPEAK FOR THEMSELVES

Chapter Written by Catherine D'Ignazio and Lauren Klein[1]

Learning Objectives

- Understand the importance of context in data collection, analysis, and interpretation,

---

1. Excerpt from the book Data Feminism, Creative Commons Attribution 4.0 International License (CC-BY 4.0). It has been modified to include learning outcomes, key takeaways, and exercises.

recognizing how it can either reinforce or challenge existing power structures.

- Explain the ethical considerations in data science, including the need to avoid deficit narratives and to be transparent about data limitations.
- Gain insights into emerging practices for providing context to data, such as data biographies, datasheets for datasets, and data user guides.

PRINCIPLE: CONSIDER CONTEXT

> Data feminism asserts that data are not neutral or objective. They are the products of unequal social relations, and this context is essential for conducting accurate, ethical analysis.

In April 2014, 276 young women were kidnapped from their high school in the town of Chibok in northern Nigeria. Boko Haram, a militant terrorist group, claimed responsibility for the attacks. The press coverage, both in Nigeria and around the world, was fast and furious. SaharaReporters.com challenged the government's ability to keep its students safe. CNN covered parents' anguish. The *Japan Times* connected the kidnappings to the increasing unrest in Nigeria's northern states. And the BBC told the story of a girl who had managed to evade the kidnappers. Several weeks after this initial reporting, the popular blog *FiveThirtyEight* published its own data-driven story about the event, titled "Kidnapping of Girls in Nigeria Is Part of a Worsening Problem."[1] The story

1. See Mona Chalabi, "Kidnapping of Girls in Nigeria Is Part of a Worsening

reported skyrocketing rates of kidnappings. It asserted that in 2013 alone there had been more than 3,608 kidnappings of young women. Charts and maps accompanied the story to visually make the case that abduction was at an all-time high (figure 6.1).

Shortly thereafter, the news website had to issue an apologetic retraction because its numbers were just plain wrong. The outlet had used the Global Database of Events, Language and Tone (GDELT) as its data source. GDELT is a big data project led by computational social scientist Kalev Leetaru. It collects news reports about events around the world and parses the news reports for actors, events, and geography with the aim of providing a comprehensive set of data for researchers, governments, and civil society. GDELT tries to focus on conflict—for example, whether conflict is likely between two countries or whether unrest is sparking a civil war—by analyzing media reports. However, as political scientist Erin Simpson pointed out to *FiveThirtyEight* in a widely cited Twitter thread, GDELT's primary data source is *media reports* (figure 6.2).[2] The project is not at a stage

Problem," *FiveThirtyEight*, May 8, 2014, https://fivethirtyeight.com/features/nigeria-kidnapping/.

2. You can see the whole thread on the archived version of Storify at https://web.archive.org/web/20140528062637/https://storify.com/AthertonKD/if-a-data-point-has-no-context-does-it-have-any-me, as well as on

at which its data can be used to make reliable claims about *independent cases* of kidnapping. The kidnapping of schoolgirls in Nigeria was a single event. There were thousands of global media stories about it. Although GDELT de-duplicated some of those stories to a single event, it still logged, erroneously, that hundreds of kidnapping events had happened that day. The *FiveThirtyEight* report had counted each of those GDELT pseudoevents as a separate kidnapping incident.

Simpson's account directly: Erin Simpson (@charlie_simpson), "So if #GDELT says there were 649 kidnappings in Nigeria in 4 months, WHAT IT'S REALLY SAYING is there were 649 news stories abt kidnappings," Twitter, May 13, 2014, 4:04 p.m., https://twitter.com/charlie_simpson/status/466308105416884225.

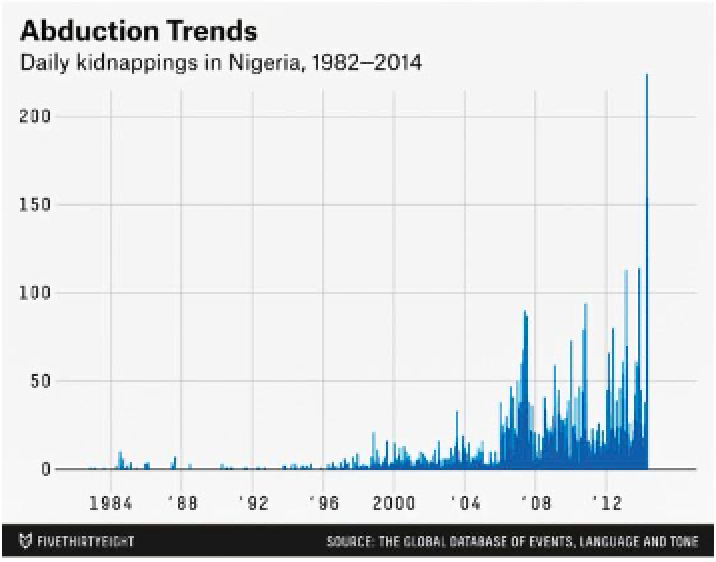


Figure 1: In 2014, FiveThirtyEight erroneously charted counts of "daily kidnappings" in Nigeria. The news site failed to recognize that the data source it was using was not counting events, but rather media reports about events. Or some events and some media reports. Or it was counting something, but we are still not sure what.

The error was embarrassing for *FiveThirtyEight*, not to mention for the reporter, but it also helps to illustrate some of the larger problems related to data found "in the wild." First, the hype around "big data" leads to projects like GDELT wildly overstating the completeness and accuracy of its data and algorithms. On the website and in publications, the project leads have stated that GDELT is "an initiative to construct a catalog of human societal-scale behavior and beliefs across all countries of the world, connecting every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what's happening around the world, what its context is and who's involved, and how the world is feeling about it, every single day."[3] That giant mouthful describes no small or impotent big data tool. It is clearly Big Dick Data.

Figure 2: Two tweets by Erin Simpson in response to FiveThirtyEight's erroneous interpretation of the GDELT dataset. Tweets by Erin Simpson on May 13, 2014.

3. Kalev Leetaru, "The GDELT Project," GDELT, accessed May 12, 2018, https://www.gdeltproject.org/.

Big Dick Data is a formal, academic term that we, the authors, have coined to denote big data projects that are characterized by patriarchal, cis-masculinist, totalizing fantasies of world domination as enacted through data capture and analysis. Big Dick Data projects ignore context, fetishize size, and inflate their technical and scientific capabilities.[4] In GDELT's case, the question is whether we should take its claims of big data at face value or whether the Big Dick Data is trying to trick funding organizations into giving the project massive amounts of research funding. (We have seen this trick work many times before.)

The GDELT technical documentation does not provide any more clarity as to whether it is counting media reports (as Simpson asserts) or single events. The database

4. We would like to make clear that our association of Big Dick Data with cis-masculinism is intended to call attention to how cis-heteropatriarchy currently works to dominate data studies, and not to offend or obscure the experiences of trans, gender non-conforming, non-binary, or intersex people. On the range of expressions of masculinity and the dicks that accompany them, see Amanda Phillips, "Dicks Dicks Dicks: Hardness and Flaccidity in (Virtual Masculinity)," *Flow: A Critical Forum on Media and Culture*, March 23, 2017, https://www.flowjournal.org/2017/11/dicks-dicks-dicks/. As an example of a critique of big data that does not rely upon the dick as signifier, see Jen Jack Gieseking, "Size Matters to Lesbians, Too: Queer Feminist Interventions into the Scale of Big Data," *Professional Geographer* 70, no. 1 (2018): 150–156. We thank the members of the Data Feminism reading group for their feedback on this term, which has pushed us to clarify our commitments.

FiveThirtyEight used is called the GDELT Event Database, which certainly makes it sound like it's counting events. The GDELT documentation states that "if an event has been seen before it will not be included again," which also makes it sound like it's counting events. And a 2013 research paper related to the project confirms that GDELT is indeed counting events, but only events that are unique to specific publications. So it's counting events, but with an asterisk. Compounding the matter, the documentation offers no guidance as to what kinds of research questions are appropriate to ask the database or what the limitations might be. People like Simpson who are familiar with the area of research known as *event detection*, or members of the GDELT community, may know to not believe (1) the title of the database, (2) the documentation, and (3) the marketing hype. But how would outsiders, let alone newcomers to the platform, ever know that?

We've singled out GDELT, but the truth is that it's not very different from any number of other data repositories out there on the web. There are a proliferating number of portals, observatories, and websites that make it possible to download all manner of government, corporate, and scientific data. There are APIs that make it possible to write little programs to query massive datasets (like, for instance, all of Twitter) and download them in a structured way.[5] There are test datasets for

5. APIs allow a little program one writes to talk to other computers over the internet

network analysis, machine learning, social media, and image recognition. There are fun datasets, curious datasets, and newsletters that inform readers of datasets to explore for journalism or analysis.[6] In our current moment, we tend to think of this unfettered access to information as an inherent good. And in many ways, it *is* kind of amazing that one can just google and download data on, for instance, pigeon racing, the length of guinea pig teeth, or every single person accused of witchcraft in Scotland between 1562 and 1736—not to mention truckloads and truckloads of tweets.[7]

---

that are ready to receive data queries. Twitter, Zillow, and MOMA are some examples of large entities that have APIs available to programatically download data.

6. Here are some of our favorites: Dogs of Zurich (https://www.europeandataportal.eu/data/en/dataset/https-data-stadt-zuerich-ch-dataset-pd_stapo_hundenamen); UFO sightings (https://www.kaggle.com/NUFORC/ufo-sightings); all of the cartoon-based murals of Brussels (https://opendata.brussels.be/explore/dataset/comic-book-route/images/); Things Lost on the New York City subway system (http://advisory.mtanyct.info/LPUWebServices/CurrentLostProperty.aspx); and a list of abandoned shopping carts in Bristol (https://data.gov.uk/dataset/abandoned-shopping-trolleys-bristol-rivers). Some of the best of the newsletters include *Data Is Plural*, curated by Jeremy Singer-Vine, who is the data editor for Buzzfeed; and *Numlock News*, a daily email newsletter by Walt Hickey, which tries to provide some context around the numbers we see in the news.

7. "Scottish Witchcraft," Data.world, May 18, 2017, https://data.world/history/scottish-witchcraft.

And though the schooling on data verification received by *FiveThirtyEight* was rightly deserved, there is a much larger issue that remains unaddressed: the issue of context. As we've discussed throughout this book, one of the central tenets of feminist thinking is that all knowledge is *situated*. A less academic way to put this is that *context matters*. When approaching any new source of knowledge, whether it be a dataset or dinner menu (or a dataset of dinner menus), it's essential to ask questions about the social, cultural, historical, institutional, and material conditions under which that knowledge was produced, as well as about the identities of the people who created it.[8] Rather than seeing knowledge artifacts, like datasets, as raw input that can be simply fed into a statistical analysis or data visualization, a feminist approach insists on connecting data back to the context in which they were produced. This context allows us, as data scientists, to better understand any functional limitations of the data and any associated ethical obligations, as well as how the power and privilege that contributed to their making may be obscuring the truth.

Situating Data on the Wild

8. Trevor Muñoz and Katie Rawson, "Data Dictionary," Curating Menus, 2016, accessed April 23, 2019, http://curatingmenus.org/data_dictionary/.

Wild Web

The major issue with much of the data that can be downloaded from web portals or through APIs is that they come without context or metadata. If you are lucky you *might* get a paragraph about where the data are from or a data dictionary that describes what each column in a particular spreadsheet means. But more often than not, you get something that looks like figure 6.3.

The data shown in the figure—open budget data about government procurement in São Paulo, Brazil—do not look very technically complicated. The complicated part is figuring out how the business process behind them works. How does the government run the bidding process? How does it decide who gets awarded a contract? Are all the bids published here, or just the ones that were awarded contracts? What do terms like *competition*, *cooperation agreement*, and *terms of collaboration* mean to the data publisher? Why is there such variation in the publication numbering scheme? These are only a few of the questions one might ask when first encountering this dataset. But without answers to even some of these questions—to say nothing of the local knowledge required to understand how power is operating in this particular ecosystem—it would be difficult to even begin a data exploration or analysis project.

This scenario is not uncommon. Most data arrive on our computational doorstep context-free. And this lack of context becomes even more of a liability when accompanied by the kind of marketing hype we see in GDELT and other Big Dick Data projects. In fact, the 1980s version of these claims is what led Donna Haraway to propose the concept of situated knowledge in the first place.[9] Subsequent feminist work has drawn on the concept of situated knowledge to elaborate ideas about ethics and responsibility in relation to knowledge-making.[10] Along this line of thinking, it becomes the responsibility of the person evaluating that knowledge, or building upon it, to ensure that its "situatedness" is taken into account. For example, information studies scholar Christine Borgman advocates for understanding data in relation to the "knowledge infrastructure" from which they originate. As Borgman defines it, a *knowledge infrastructure* is "an ecology of people, practices, technologies, institutions, material

---

9. Haraway uses the phrase "unlocatable, and so irresponsible, knowledge claims." Donna Haraway, "Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective," *Feminist Studies* 14, no. 3 (Autumn 1988): 575–599, https://doi.org/10.2307/3178066.

10. For example, philosopher Lorraine Code argues that connecting knowledge to its specific biographic, historical, and geographic locations leads to "more responsible knowings." Code, *Ecological Thinking: The Politics of Epistemic Location* (New York: Oxford University Press, 2006).

objects, and relationships."[11] In short, it is the context that makes the data possible.

11. Christine L. Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World* (Cambridge, MA: MIT Press, 2015).

Nr. Publicação	Licitador	Modalidade	Dt. Abertura	Objeto
01-PREF/SECOM/2018	Secretaria do Governo Municipal - SGM	CONCORRÊNCIA	10/06/2019 14:00	Contratação de empresa para prestação de serviços de assessoria de imprensa e comunicação para a PREF/SECOM
03/SGM-2019	SGM - Administração de Compras e Contratos	CONCORRÊNCIA	03/06/2019 10:30	ALIENAÇÃO DO IMÓVEL MUNICIPAL SITUADO NA AVENIDA PROFESSOR ALCEU MAYNARD ARAÚJO, NO DISTRITO DE SANTO AMARO.
01/SMPED/2019	Secretaria Municipal da Pessoa com Deficiência - SMPED	TOMADA DE PREÇOS	31/05/2019 10:30	Contratação de empresa especializada em produção e atualização de material didático orientador e informativo com produção de conteúdo em versão digital acessível, visando a subsidiar a capacitação do público alvo dos cursos e eventos oferecidos pela Secretaria Municipal da Pessoa com Deficiência - SMPED.
18/SME/2019	Secretaria Municipal de Educação - SME	PREGÃO ELETRÔNICO	30/05/2019 10:30	Registro de preços para aquisição de alimentos não perecíveis açúcar refinado.
007/2019	São Paulo Transporte S/A	PREGÃO ELETRÔNICO	27/05/2019 10:00	OBJETO: AQUISIÇÃO DE 6 (SEIS) EQUIPAMENTOS APPLIANCE DO TIPO UTM?s, COM LICENÇAS DE SEGURANÇA, INSTALAÇÃO E SUPORTE TÉCNICO, PELO PERÍODO DE 24 (VINTE E QUATRO) MESES
109/SMADS/2019	Secretaria Municipal de Assistência e Desenvolvimento Social - SMADS	TERMO DE COLABORAÇÃO - EDITAL	24/05/2019 10:00	C J
108/SMADS/2019	Secretaria Municipal de Assistência e Desenvolvimento Social - SMADS	TERMO DE COLABORAÇÃO - EDITAL	24/05/2019 10:00	Centro de Acolhida com Inserção Produtiva para Adultos em Situação de Rua
001/2018/SEHAB	Secretaria Municipal de Habitação - SEHAB - GABINETE	CONCORRÊNCIA	24/05/2019 10:00	EXECUÇÃO DE OBRAS DE CONSTRUÇÃO DE EMPREENDIMENTO HABITACIONAL DE INTERESSE SOCIAL E DE USO MISTO, DENOMINADO COLISEU, NO ÂMBITO DA OPERAÇÃO URBANA CONSORCIADA FARIA LIMA
002/SVMA/2019	Secretaria Municipal do Verde e Meio Ambiente - SVMA	CONCORRÊNCIA	23/05/2019 10:30	CONTRATAÇÃO DE SERVIÇOS TÉCNICOS ESPECIALIZADOS PARA A ELABORAÇÃO DO PLANO DE MANEJO DA ÁREA DE PROTEÇÃO AMBIENTAL (APA) BORORÉ-COLÔNIA
070/18	São Paulo Turismo - SPTURIS	PREGÃO ELETRÔNICO	22/05/2019 10:00	Contratação de empresa, sob o regime de empreitada por preço unitário, para prestação de serviços de BOMBEIRO PROFISSIONAL CIVIL, por um período de 12 (doze) meses, prorrogáveis por iguais ou menores períodos, conforme bases, especificações e condições do Edital e seus Anexos.
093/2019-SMS.G	Secretaria Municipal de Saúde - SMS	PREGÃO ELETRÔNICO	22/05/2019 09:00	Registro de preços para o fornecimento de PAPEL CREPADO E SWAB, ALCOOL 70% PARA ANTI-SEPSIA.
121/2019-SMS.G	Secretaria Municipal de Saúde - SMS	PREGÃO ELETRÔNICO	21/05/2019 10:30	Registro de preços para o fornecimento de KIT PARA IDENTIFICAÇÃO QUALITATIVA PARA O COMPLEXO M. TUBERCULOSIS.
16/SME/2019	Secretaria Municipal de Educação - SME	PREGÃO ELETRÔNICO	21/05/2019 10:30	Registro de preço para aquisição de Item A: Sardinha em óleo comestível e Item B: Atum em pedaços em conserva.
166/2019	Autarquia Hospitalar Municipal - AHM	PREGÃO ELETRÔNICO	21/05/2019 09:30	AQUISIÇÃO DE SULFAMETOXAZOL 80 MG/ML + TRIMETOPRIMA 16 MG/ML 5 ML, PARA AS UNIDADES DA AUTARQUIA HOSPITALAR MUNICIPAL.
119/2019-SMS.G	Secretaria Municipal de Saúde - SMS	PREGÃO ELETRÔNICO	20/05/2019 10:30	Registro de preços para o fornecimento de ETIQUETA TÉRMICA CONTÍNUA, AUTOADESIVA, PARA IMPRESSAO TÉRMICA ? 62MM X 15M.
117/2019-SMS.G	Secretaria Municipal de Saúde - SMS	PREGÃO ELETRÔNICO	20/05/2019 09:30	Aquisição de MATERIAL ODONTOLÓGICO - FÓRCEPS PARA USO ODONTOLÓGICO.
047/2019-HMEC	Hospital Municipal Maternidade-Escola Dr. Mario de Moraes Altenfelder Silva	PREGÃO ELETRÔNICO	20/05/2019 09:00	BERACTANTO SUSPENSÃO INTRA-TRAQUEAL 25 MG/ML FAM 8,0 ML ? FAM
103/2019-SMS.G	Secretaria Municipal de Saúde - SMS	PREGÃO ELETRÔNICO	17/05/2019 10:30	Registro de preços para o fornecimento de MATERIAL DE LABORATÓRIO - COLETOR UNIVERSAL ESTÉRIL, PIPETA DE TRANSFERÊNCIA E SWAB DE RAYON.
035/2019-HMEC	Hospital Municipal Maternidade-Escola Dr. Mario de Moraes Altenfelder Silva	PREGÃO ELETRÔNICO	17/05/2019 10:00	PLACA DESCARTAVEL PARA ELETROCIRURGIA
002/2019	São Paulo Obras - SP Obras	TOMADA DE PREÇOS	17/05/2019 09:30	Contratação de empresa especializada em engenharia e arquitetura para execução das obras de reforma para implantação do DESCOMPLICA SP ? UNIDADE SÃO MATEUS .

Figure 3: Open budget data about procurement and expenses from the São Paulo prefecture in Brazil. Although Brazil has some of the most progressive transparency laws on the books, the data that are published aren't necessarily always accessible or usable by citizens and residents. In 2013, researcher Gisele Craveiro worked with civil society

organizations to give this open budget data more context. Images from SIGRC for the Prefecture of São Paulo, Brazil.

Ironically, some of the most admirable aims and actions of the open data movement have worked against the ethical urgency of providing context, however inadvertently. *Open data* describes the idea that anyone can freely access, use, modify, and share data for any purpose. The open data movement is a loose network of organizations, governments, and individuals. It has been active in some form since the mid-2000s, when groups like the Open Knowledge Institute were founded and campaigns like Free Our Data from the *Guardian* originated to petition governments for free access to public records.[12] The goals are good ones in theory: economic development by building apps and services on open data; faster scientific

12. "Open Knowledge International." Open Knowledge International. Accessed March 27, 2019. https://okfn.org/. The *Guardian* newspaper, out of the United Kingdom, launched the Free Our Data campaign in 2006 to petition government agencies to make public data available to taxpayers and companies for free. Among other things, they focused on geographic data collected by the Royal Ordnance Survey which had restrictive licenses on reuse by citizens. The campaign was largely successful: in 2010, the United Kingdom created the Open Government License and launched data.gov.uk, one of the first national data portals in the world. See Charles Arthur and Michael Cross, "Give Us Back Our Crown Jewels," *Guardian*, March 9, 2006, https://www.theguardian.com/technology/2006/mar/09/education.epublic.

progress when researchers share knowledge; and greater transparency for journalists, citizens, and residents to be able to use public information to hold governments accountable. This final goal was a major part of the framing of former US president Obama's well-known memorandum on transparency and open government.[13] On his very first day in office, Obama signed a memorandum that directed government agencies to make all data open by default.[14] Many more countries, states, and cities have followed suit by developing open data portals and writing open data into policy. As of 2019, seventeen countries and over fifty cities and states have adopted the International Open Data Charter,

13. See Peter R. Orszag, "Memorandum for the Heads of Executive Departments and Agencies Re: Open Government Directive," Washington, DC, Executive Office of the President, December 8, 2009, https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/memoranda_2010/m10-06.pdf.

14. Although the movement under Obama was toward openness (Orszag, "Memorandum for the Heads of Executive Departments and Agencies Re: Open Government Directive"), the current administration has retreated from this position, according to a Sunlight Foundation audit, which found that "the Open Government Initiative, Open Government Partnership, and related programs, initiatives and partnerships across the federal government are being ignored, neglected or even forgotten in federal agencies." Briana Williams, "Under Trump, U.S. Government Moves from /Open to /Closed," Sunlight Foundation, January 24, 2018, https://sunlightfoundation.com/2018/01/24/under-trump-u-s-government-moves-from-open-to-closed/.

which outlines a set of six principles guiding the publication and accessibility of government data.[15]

In practice, however, limited public funding for technological infrastructure has meant that governments have prioritized the "opening up" part of open data—publishing spreadsheets of things like license applications, arrest records, and flood zones—but lack the capacity to provide any context about the data's provenance, let alone documentation that would allow the data to be made accessible and usable by the general public. As scholar Tim Davies notes, raw data dumps might be good for starting a conversation, but they cannot ensure engagement or accountability.[16] The reality is that many published datasets sit idle on their portals, awaiting users to undertake the intensive work of deciphering the bureaucratic arcana that obscures their significance. This phenomenon has been called *zombie data*: datasets that have been published without any purpose or clear use case in mind.[17]

---

15. "The International Open Data Charter," Open Data Charter, accessed March 27, 2019, https://opendatacharter.net/principles/.

16. Tim Davies, "Exploring Participatory Public Data Infrastructure in Plymouth," *Public Sector Blogs*, September 11, 2017, https://www.publicsectorblogs.org.uk/2017/09/exploring-participatory-public-data-infrastructure-in-plymouth-tim-davies/.

17. *Zombie data* was named by Daniel Kaufmann, an economist with the Revenue

Zombies might be bad for brains, but is zombie data really a problem? *Wired* magazine editor Chris Anderson would say, emphatically, "No." In a 2008 *Wired* article, "The End of Theory," Anderson made the now-infamous claim that "the numbers speak for themselves."[18] His main assertion was that the advent of big data would soon allow data scientists to conduct analyses at the scale of the entire human population, without needing to restrict their analysis to a smaller sample. To understand his claim, you need to understand one of the basic premises of statistics.

Statistical inference is based on the idea of sampling: that

Watch Institute. Joel Gurin, "Open Governments, Open Data: A New Lever for Transparency, Citizen Engagement, and Economic Growth," *SAIS Review of International Affairs* 34, no. 1 (Winter 2014): 71–82. While the name is certainly evocative, it's also important to acknowledge the history of zombies, which can be traced to seventeenth-century Haiti as a response to the incursion of slavery. As Mike Mariani helpfully summarizes, enslaved Haitians "believed that dying would release them back to *lan guinée*, literally Guinea, or Africa in general, a kind of afterlife where they could be free." But "those who took their own lives wouldn't be allowed to return to *lan guinée*. Instead, they'd be condemned to skulk the Hispaniola plantations for eternity, an undead slave at once denied their own bodies and yet trapped inside them—a soulless zombie." See Mariani, "The Tragic, Forgotten History of Zombies," *Atlantic*, October 28, 2015, https://www.theatlantic.com/entertainment/archive/2015/10/how-america-erased-the-tragic-history-of-the-zombie/412264/.

18. See Chris Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," *Wired*, June 23, 2008, https://www.wired.com/2008/06/pb-theory/.

you can infer things about a population (or other large-scale phenomenon) by studying a random and/or representative sample and then mapping those findings back on the population (or phenomenon) as a whole. Say that you want to know who all of the 323 million people in the US will vote for in the coming presidential election. You couldn't contact all of them, of course, but you could call three thousand of them on the phone and then use those results to predict how the rest of the people would likely vote. There would also need to be some statistical modeling and theory involved, because how do you know that those three thousand people are an accurate representation of the whole population? This is where Anderson made his intervention: at the point at which we have data collected on the entire population, we no longer need modeling, or any other "theory" to first test and then prove. We can look directly at the data themselves.

Now, you can't write an article claiming that the basic structure of scientific inquiry is obsolete and not expect some pushback. Anderson wrote the piece to be provocative, and sure enough, it prompted numerous responses and debates, including those that challenge the idea that this argument is a "new" way of thinking in the first place (e.g., in the early seventeenth century, Francis Bacon argued for a form of inductive reasoning, in which the scientist gathers data, analyzes them, and only thereafter forms a hypothesis).[19] One of Anderson's major examples is Google Search. Google's search algorithms don't need to have a hypothesis about *why* some websites have more incoming links—other pages that link to the site—than others; they just need a way to determine the number of links so they can use that number to determine the popularity and relevance of the site in search results. We no longer need causation, Anderson insists: "Correlation is enough."[20] But what happens when the number of links is also highly correlated with sexist, racist, and pornographic results?

The influence of racism, sexism, and colonialism is precisely what we see described in *Algorithms of Oppression*,

19. Fulvio Mazzocchi makes the connection between Bacon and big data in "Could Big Data Be the End of Theory in Science?," EMBO reports 16, no. 10 (2015): 1250–1255. While Bacon's *Novum Organum* (1620) was indeed a masterful work that influenced centuries of scientists, he was not alone in his promulgation of a (proto) scientific method. Margaret Cavendish (1623–1717), for example, was an author of both natural philosophy (as scientific theory was known at the time) and science fiction. In fact, her scientific treatise, *Observations upon Experimental Philosophy*, was published alongside her science fiction text, *The Blazing World* (1666), and together they worked to challenge the domination of science by men—a reality even in the seventeenth century.

20. Historian Matthew Jones has written an intellectual history of this line of thinking and demonstrates how it has led to a computational "culture of predictive utility" in which prediction is prized above other possible measures of success. See Jones, "How We Became Instrumentalists (Again)," *Historical Studies in the Natural Sciences* 48, no. 5 (November 5, 2018): 673–684.

information studies scholar Safiya Umoja Noble's study of the harmful stereotypes about Black and Latinx women perpetuated by search algorithms such as Google's. As discussed in chapter 1, Noble demonstrates that Google Search results do not simply correlate with our racist, sexist, and colonialist society; that society *causes* the racist and sexist results. More than that, Google Search reinforces these oppressive views by ranking results according to how many other sites link to them. The rank order, in turn, encourages users to continue to click on those same sites. Here, correlation without context is clearly not enough because it recirculates racism and sexism and perpetuates inequality.[21]

There's another reason that context is necessary for making sense of correlation, and it has to do with how racism, sexism, and other forces of oppression enter into the environments in which data are collected. The next example has to do with sexual assault and violence. If you do not want to read about these topics, you may want to skip ahead to the next section.

In April 1986, Jeanne Clery, a student at Lehigh University, was sexually assaulted and murdered in her dorm room. Her

21. Safiya Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism* (New York: NYU Press, 2018), 80–81.

parents later found out that there had been thirty-eight violent crimes at Lehigh in the prior three years, but nobody had viewed that as important data that should be made available to parents or to the public. The Clerys mounted a campaign to improve data collection and communication efforts related to crimes on college campuses, and it was successful: the Jeanne Clery Act was passed in 1990, requiring all US colleges and universities to make on-campus crime statistics available to the public.[22]

So we have an ostensibly comprehensive national dataset about an important public topic. In 2016, three students in Catherine's data journalism class at Emerson College—Patrick Torphy, Michaela Halnon, and Jillian Meehan—downloaded the Clery Act data and began to explore it, hoping to better understand the rape culture that has become pervasive on college campuses across the United States.[23] They soon became

22. See https://clerycenter.org/policy-resources/the-clery-act/. The data include separate and specific numbers on sexual assault, dating violence, domestic violence, and stalking. It includes sexual assault incidents experienced by women, men, and nonbinary people.

23. The term *rape culture* was coined by second-wave feminists in the 1970s to denote a society in which male sexual violence is normalized and pervasive, victims are blamed, and the media exacerbates the problem. Rape culture includes jokes, music, advertising, laws, words, and images that normalize sexual violence. In 2017, following the election of a US president who joked about sexual assault on the campaign trail and the exposé of Harvey Weinstein's predatory behavior in

puzzled, however. Williams College, a small, wealthy liberal arts college in rural Massachusetts, seemed to have an epidemic of sexual assault, whereas Boston University (BU), a large research institution in the center of the city, seemed to have strikingly few cases relative to its size and population (not to mention that several high-profile sexual assault cases at BU had made the news in recent years).[24] The students were suspicious of these numbers, and investigated further. After comparing the Clery Act data with anonymous campus climate surveys (figure 6.4), consulting with experts, and interviewing survivors, they discovered, paradoxically, that the truth was closer to the *reverse* of the picture that the Clery Act data suggest. Many of the colleges with higher reported rates of sexual assault were actually places where more institutional resources were being devoted to support for survivors.[25]

Hollywood, high-profile women began speaking out against rape culture with the #MeToo hashtag. #MeToo, a movement started over a decade ago by activist Tarana Burke, encourages survivors to break their silence and build solidarity to end sexual violence.

24. In 2012, two members of BU's hockey team were charged with sexual assault, and a report by the university found that the team had created a "culture of sexual entitlement." See Mary Carmichael, "Graphic Details Emerge from BU Hockey Panel Reports," *Boston Globe*, September 6, 2012, https://www.boston.com/news/local-news/2012/09/06/graphic-details-emerge-from-bu-hockey-panel-reports.

25. The students' full story is excellent. You can read it here: Patrick Torphy, Michaela Halnon, and Jillian Meehan, "Reporting Sexual Assault: What the Clery Act

As for the colleges with lower numbers, this is also explained by context. The Clery Act requires colleges and universities to provide annual reports of sexual assault and other campus crimes, and there are stiff financial penalties for not reporting. But the numbers are self-reported, and there are also strong financial incentives for colleges *not* to report.[26] No college wants to tell the government—let alone parents of prospective students—that it has a high rate of sexual assault on campus. This is compounded by the fact that survivors of sexual assault often do not want to come forward—because of social stigma, the trauma of reliving their experience, or the resulting lack of social and psychological support. Mainstream culture has taught survivors that their experiences will not be treated with

---

Doesn't Tell Us," *Atavist*, April 26, 2016, https://cleryactfallsshort.atavist.com/reporting-sexual-assault-what-the-clery-act-doesnt-tell-us.

26. Sixteen staff members at the US Department of Education are devoted to monitoring the more than seven thousand higher-education institutions in the country, so it is unlikely that underreporting by an institution would be discovered, except in very high-profile cases. See Michael Stratford, "Clery Fines: Proposed vs. Actual," *Inside HigherEd*, July 17, 2014, https://www.insidehighered.com/news/2014/07/17/colleges-often-win-reduction-fines-federal-campus-safety-violations. For example, the Sandusky Case at Penn State involved systematic sexual abuse of young boys by a football coach, and the university was subsequently fined $2.4 million for failing to properly report and disclose these crimes.

care and that they may in fact face more harm, blame, and trauma if they do come forward.[27]

27. In this context, one might consider the decision of Christine Blasey Ford to testify about her assault by (now) US Supreme Court Justice Brett Kavanaugh. Coming forward involved relinquishing her privacy and reliving her trauma multiple times over, on a national stage.

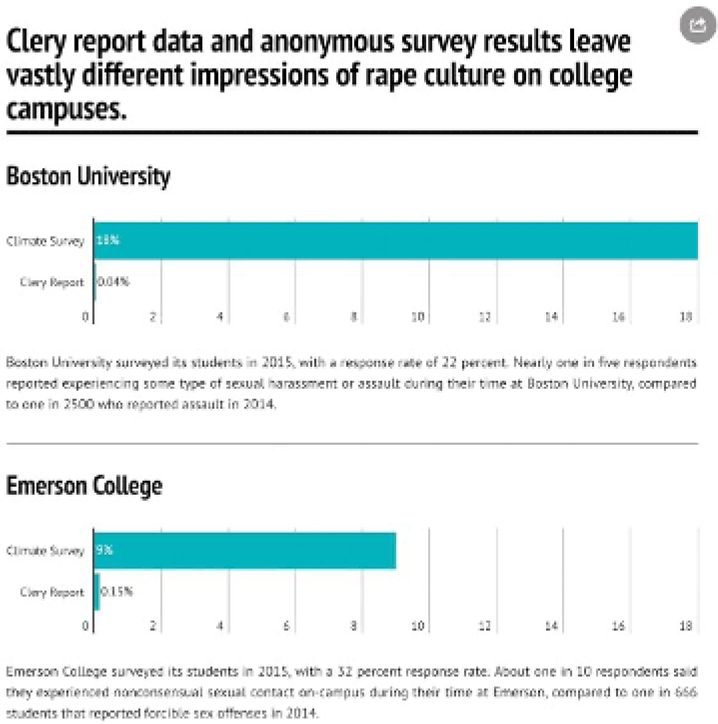


Figure 4: Data journalism students at Emerson College were skeptical of the self-reported Clery Act data and decided to compare the Clery Act results with anonymous campus climate survey results about nonconsensual sexual contact. Although there are data-quality issues with both datasets, the students assert that if institutions are providing adequate support for survivors, then there will be less of a gap between the Clery-reported data and the proportion of students that report nonconsensual sexual conduct. Courtesy of Patrick Torphy, Michaela Halnon, and Jillian Meehan, 2016.

There are further power differentials reflected in the data when race and sexuality are taken into account. For example, in 2014, twenty-three students filed a complaint against Columbia University, alleging that Columbia was systematically mishandling cases of rape and sexual violence reported by LGBTQ students. Zoe Ridolfi-Starr, the lead student named in the complaint, told the *Daily Beast*, "We see complete lack of knowledge about the specific dynamics of sexual violence in the queer community, even from people who really should be trained in those issues."[28]

Simply stated, there are imbalances of power in the *data setting*—to use the phrase coined by Yanni Loukissas that we discussed in chapter 5—so we cannot take the numbers in the dataset at face value. Lacking this understanding of power in the collection environment and letting the numbers "speak for themselves" would tell a story that is not only patently false but could also be used to reward colleges that are systematically underreporting and creating hostile environments for survivors. Deliberately undercounting cases of sexual assault leads to being rewarded for underreporting. And the silence around sexual assault continues: the administration is silent, the campus culture is silent, the dataset is silent.[29]

28. Abigail Golden, "Is Columbia University Mishandling LGBT Rape Cases?," *Daily Beast*, April 30, 2014, https://www.thedailybeast.com/is-columbia-university-mishandling-lgbt-rape-cases?ref=scroll.

29. Sara Ahmed has written powerfully on the violent effects of this silencing of assault victims. "Silence enables the reproduction of the culture of harassment and abuse.

Raw Data, Cooked Data, Cooking

As demonstrated by the Emerson College students, one of the key analytical missteps of work that lets "the numbers speak for themselves" is the premise that data are a *raw input*. But as Lisa Gitelman and Virginia Jackson have memorably explained, data enter into research projects already fully cooked—the result of a complex set of social, political, and historical circumstances. "'Raw data' is an oxymoron," they assert, just like "jumbo shrimp."[30] But there is an emerging

---

When we don't speak about violence we reproduce violence. Silence about violence is violence," she explains. Ahmed, "Speaking Out," *Feministkilljoys* (blog), June 2, 2016, https://feministkilljoys.com/2016/06/02/speaking-out/.

30. Lisa Gitelman and Virginia Jackson, "Introduction," in *"Raw Data" Is an Oxymoron* (Cambridge, MA: MIT Press, 2013), 2. Here they are following a statement from information studies scholar Geoffrey Bowker, "Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care," as quoted in *Memory Practices in the Sciences* (Cambridge, MA: MIT Press, 2005). The dichotomy between "raw" and "cooked," in turn, owes its source to the renowned structural anthropologist Claude Levi-Strauss. His famous book, *The Raw and the Cooked* (1964), analogizes the process of transforming nature into culture as akin to the process of transforming raw food into cooked. Your false binary and hidden hierarchy alarm bells should already be going off; and indeed, much of the work of the feminist theory of the early 1970s was to challenge this false dichotomy, as well as the assumptions (and examples) that it rested upon. See Lévi-

class of "data creatives" whose very existence is premised on their ability to *context-hop*—that is, their ability to creatively mine and combine data to produce new insights, as well as work across diverse domains. This group includes data scientists, data journalists, data artists and designers, researchers, and entrepreneurs—in short, pretty much everyone who works with data right now. They are the strangers in the dataset that we spoke of in chapter 5.

Data's new creative class is highly rewarded for producing work that creates new value and insight from mining and combining conceptually unrelated datasets. Examples include Google's now defunct Flu Trends project, which tried to geographically link people's web searches for flu symptoms to actual incidences of flu.[31] Or a project of the *Sun Sentinel* newspaper, in Fort Lauderdale, Florida, which combined police license plate data with electronic toll records to prove that cops were systematically and dangerously speeding on Florida highways.[32] Sometimes these acts of creative synthesis work out well; the *Sun Sentinel* won a Pulitzer for its reporting

---

Strauss, trans. John Weightman and Doreen Weightman, *The Raw and the Cooked: Introduction to a Science of Mythology*, vol. 1 (New York: Harper & Row, 1969).

31. "Google Flu Trends," accessed August 6, 2019, https://www.google.org/flutrends/about/.

32. Sally Kestin and John Maines, "Cops among Florida's Worst Speeders, Sun Sentinel Investigation Finds," *Sun Sentinel*, February 11, 2012.

and a number of the speeding cops were fired. But sometimes the results are not quite as straightforward. Google Flu Trends worked well until it didn't, and subsequent research has shown that Google searches cannot be used as 1:1 signals for actual flu phenomena because they are susceptible to external factors, such as what the media is reporting about the flu.[33]

Instead of taking data at face value and looking toward future insights, data scientists can first interrogate the context, limitations, and validity of the data under use. In other words, one feminist strategy for considering context is to consider the cooking process that produces "raw" data. As one example, computational social scientists Derek Ruths and Jürgen Pfeffer write about the limitations of using social media data for behavioral insights: Instagram data skews young because Instagram does; Reddit data contains far more comments by men than by women because Reddit's overall membership is majority men. They further show how research data acquired

33. A brilliant idea—to try to link searches for flu symptoms to actual cases of the flu to see if one could predict where the next outbreak would be—Google Flu Trends seemed to work for the first couple years. Then, in the 2012–2013 flu season, Google estimated more than double the flu cases that the CDC did. This discrepancy was possibly due to media panic about swine flu, to Google updating its technology to include recommendations, or perhaps to something else. These are the dangers of prioritizing prediction and utility over causation and context: it all works temporarily, until something in the environment changes. See David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani, "The Parable of Google Flu: Traps in Big Data Analysis," *Science* 343, no. 6176 (2014): 1203–1205.

from those sources are shaped by sampling because companies like Reddit and Instagram employ proprietary methods to deliver their data to researchers, and those methods are never disclosed.[34] Related research by Devin Gaffney and J. Nathan Matias took on a popular corpus that claimed to contain "every publicly available Reddit comment."[35] Their work showed the that the supposedly complete corpus is missing at least thirty-six million comments and twenty-eight million submissions.

Exploring and analyzing what is missing from a dataset is a powerful way to gain insight into the cooking process—of both the data and of the phenomenon it purports to represent. In some of Lauren's historical work, she looks at actual cooks as they are recorded (or not) in a corpus of thirty thousand

---

34. In the paper "Tampering with Twitter's Sample API," Jürgen Pfeffer, Katja Mayer, and Fred Morstatter demonstrate how the opacity of sampling done by platforms makes the data vulnerable to manipulation. Pfeffer, Mayer, and Morstatter, "Tampering with Twitter's Sample API," *EPJ Data Science* 7, no. 1 (December 19, 2018).

35. Gaffney and Matias found that the supposedly complete corpus is missing at least thirty-six million comments and twenty-eight million submissions. At least fifteen peer-reviewed studies have used the dataset for research studies on topics like politics, online behavior, breaking news, and hate speech. Depending on what the researchers used the corpus for, the missing data may have affected the validity of their results. Devin Gaffney, and J. Nathan Matias, "Caveat Emptor, Computational Social Science: Large-Scale Missing Data in a Widely-Published Reddit Corpus," *PLOS ONE* 13, no. 7 (July 6, 2018).

letters written by Thomas Jefferson, as shown in figure 6.5.[36] Some may already know that Jefferson is considered the nation's "founding foodie."[37] But fewer know that he relied upon an enslaved kitchen staff to prepare his famous food.[38] In "The Image of Absence," Lauren used named-entity recognition, a natural language processing technique, to identify the places in Jefferson's personal correspondence where he named these people and then used social network analysis to approximate the extent of the relationships among them. The result is a visual representation of all of the work that Jefferson's enslaved staff put into preparing his meals but that he did not acknowledge—at least not directly—in the text of the letters themselves.

36. Lauren F. Klein, "The Image of Absence: Archival Silence, Data Visualization, and James Hemings," *American Literature* 85, no. 4 (Winter 2013): 661–688.

37. The title of a book by Dave Dewitt, *The Founding Foodies: How Washington, Jefferson, and Franklin Revolutionized American Cuisine* (Naperville, IL: Sourcebooks, 2010).

38. The subject of Adrian Miller's *The President's Kitchen Cabinet: The Story of the African Americans Who Have Fed Our First Families, from the Washingtons to the Obamas* (Chapel Hill: University of North Carolina Press, 2017) and of Lauren's more academic book on the subject, *An Archive of Taste: Race and Eating in the Early United States* (Minneapolis: University of Minnesota Press, 2020).

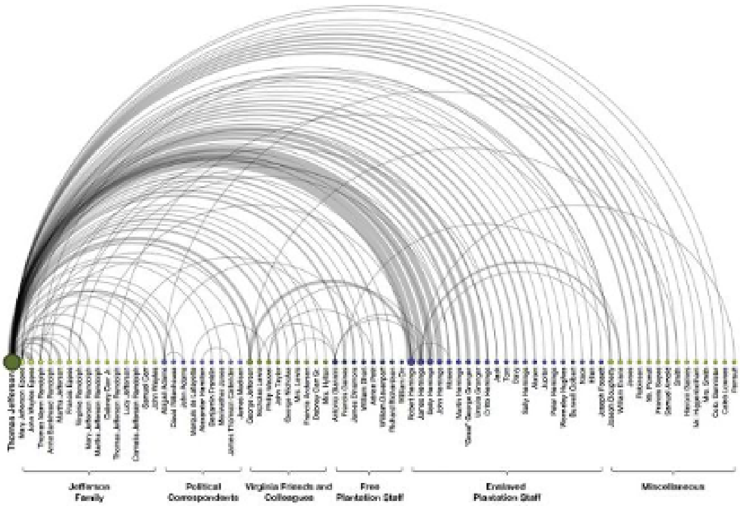


Figure 5: In "The Image of Absence" (2013), Lauren used machine learning techniques to identify the names of the people whom Thomas Jefferson mentioned in his personal correspondence and then visualized the relationships among them. The result demonstrates all of the work that his enslaved staff put into preparing Jefferson's meals but that was not directly acknowledged by Jefferson himself. Visualization by Lauren F. Klein.

On an even larger scale, computer scientists and historians at Stanford University used word embeddings—another machine learning technique—to explore gender and ethnic stereotypes across the span of the twentieth century.[39] Using

39. See Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou, "Word

several large datasets derived from sources such as the Google Books and the *New York Times*, the team showed how words like *intelligent*, *logical*, and *thoughtful* were strongly associated with men until the 1960s. Since that time, however, those words have steadily increased in association with women. The team attributed this phenomenon to the "women's movement in the 1960s and 1970s," making their work an interesting example of an attempt to quantify the impact of social movements. The paper is also notable for openly acknowledging how their methods, which involved looking at the adjectives surrounding the words *man* and *woman*, limited the scope of their analysis to the gender binary. Furthermore, the researchers did not try to assert that the data represent how women and men "are," nor did they try to "remove the bias" so that they could develop "unbiased" applications in other domains. They saw the data as what they are—cultural indicators of the changing face of patriarchy and racism—and interrogated them as such.

So, how do we produce more work like this—work that understands data as already "cooked" and then uses that data to expose structural bias? Unfortunately for Chris Anderson, the answer is that we need more theory, not less. Without theory, survey designers and data analysts must rely on their

Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes," *Proceedings of the National Academy of Sciences* 115, no. 16 (2018): E3635–E3644.

intuition, supported by "common sense" ideas about the things they are measuring and modeling. This reliance on "common sense" leads directly down the path to bias. Take the case of GDELT. Decades of research has demonstrated that events covered by the media are selected, framed, and shaped by what are called "news values": values that confirm existing images and ideologies.[40] So what is it really that GDELT is measuring? What events are happening in the world, or what the major international news organizations are focusing their attention on? The latter might be the most powerful story embedded in the GDELT database. But it requires deep context and framing to draw it out.

Refusing to acknowledge context is a power play to avoid power. It's a way to assert authoritativeness and mastery without being required to address the complexity of what the data actually represent: the political economy of the news in the case of GDELT, entrenched gender hierarchies and flawed reporting environments in the case of the Clery data, and so on. But deep context and computation are not incompatible. For example, SAFElab, a research lab at Columbia run by scholar and social worker Desmond Patton, uses artificial

40. In 1970, Daniel Halloran and colleagues wrote, "Events will be selected for news reporting in terms of their fit or consonance with pre-existing images—the news of the event will confirm earlier ideas." James Dermot Halloran, Philip Ross Courtney Elliott, and Graham Murdock, *Demonstrations and Communication: A Case Study* (London: Penguin Books, 1970).

intelligence to examine the ways that youth of color navigate violence on and offline. He and a team of social work students use Twitter data to understand and prevent gang violence in Chicago. Their data are big, and they're also complicated in ways that are both technical and social. The team is acutely aware of the history of law enforcement agencies using technology to surveil Black people, for example, and acknowledges that law enforcement continues to do so using Twitter itself. What's more, when Patton started his research, he ran into an even more basic problem: "I didn't know what young people were saying, period."[41] This was true even though Patton himself is Black, grew up in Chicago, and worked for years in many of these same neighborhoods. "It became really clear to me that we needed to take a deeper approach to social media data in particular, so that we could really grasp culture, context and nuance, for the primary reason of not misinterpreting what's being said," he explains.[42]

Patton's approach to incorporating culture, context, and nuance took the form of direct contact with and centering the perspectives of the youth whose behaviors his group sought to study. Patton and doctoral student William Frey hired formerly gang-involved youth to work on the project as

41. Desmond Patton, interview by Catherine D'Ignazio, August 30, 2018.

42. Patton, interview by D'Ignazio.

domain experts. These experts coded and categorized a subset of the millions of tweets, then trained a team of social work students to take over the coding. The process was long and not without challenges. It required that Patton and Frey create a new "deep listening" method they call the *contextual analysis of social media* to help the student coders mitigate their own bias and get closer to the intended meaning of each tweet.[43] The step after that was to train a machine learning classifier to automatically label the tweets, so that the project could categorize all of the millions of tweets in the dataset. Says Patton, "We trained the algorithm to think like a young African American man on the south side of Chicago."[44]

This approach illustrates how context can be integrated into an artificial intelligence project, and can be done with an attention to *subjugated knowledge*. This term describes the forms of knowledge that have been pushed out of mainstream institutions and the conversations they encourage. To explain this phenomenon, Patricia Hill Collins gives the example of how Black women have historically turned to "music, literature, daily conversations, and everyday behavior" as a

---

43. This method is described in detail in their paper: William R. Frey, Desmond U. Patton, Michael B. Gaskell, and Kyle A. McGregor, "Artificial Intelligence and Inclusion: Formerly Gang-Involved Youth as Domain Experts for Analyzing Unstructured Twitter Data," *Social Science Computer Review* (2018).

44. Patton, interview by D'Ignazio.

result of being excluded from "white male-controlled social institutions."[45] These institutions include academia, or—for a recent example raised by sociologist Tressie McMillan Cottom—the op-ed section of the *New York Times*.[46] And because they circulate their knowledge in places outside of those mainstream institutions, that knowledge is not seen or recognized by those institutions: it becomes *subjugated*.

The idea of subjugated knowledge applies to other minoritized groups as well, including the Black men from Chicago whom Patton sought to understand. An approach that did not attend to this context would have resulted in significant errors. For example, a tweet like "aint kill yo mans & ion kno ya homie" would likely have been classified as aggressive or violent, reflecting its use of the word "kill." But drawing on the knowledge provided by the young Black men they hired for the project, Frey and Patton were able to show that many tweets like this one were references to song lyrics, in this case the Chicago rapper Lil Durk. In other words, these tweets are about sharing culture, not communicating threats.[47]

---

45. Collins, *Black Feminist Thought*.

46. See "Tressie McMillan Cottom—Upending Stereotypes of Black Womanhood with 'Thick,'" *The Daily Show with Trevor Noah*, video, 7:20, January 21, 2019, https://www.youtube.com/watch?v=EYNu6yvv8HU.

47. Context is crucial for understanding social media conversations. This becomes a particularly fraught problem once we start automating meaning-making with

In the case of SAFElab, as with all research projects that seek to make use of subjugated knowledge, there is also significant human, relational infrastructure required. Frey and Patton have built long-term relationships with individuals and organizations in the community they study. Indeed, Frey lives and works in the community. In addition, both Frey and Patton are trained as social workers. This is reflected in their computational work, which remains guided by the social worker's code of ethics.[48] They are using AI to broker new

---

techniques like sentiment analysis and quantitative text analysis. Language and image meanings shift and change quickly, often based on local knowledge, culture, and circumstances. Mariana Giorgetti Valente, director of the Brazilian nonprofit InternetLab, gives the example of a 2010 attack on a gay man in São Paulo in which he was hit on the head with a neon lamp. The image of a lamp then became used in hate speech online. When somebody would subsequently speak out in support of gay rights on Brazilian social media, trolls would post a lamp to communicate a threat of violence. But how would a machine-learning classifier understand that an image of a lamp is a threat without knowing this local context? Valente and InternetLab are collaborating with IT for Change in India to see how they can incorporate context into the detection of hate-speech and anti-hate-speech practices online. Mariana Valente, interview by Catherine D'Ignazio, March 11, 2019.

48. One of the principles of this code is that "social workers recognize the central importance of human relationships." As new codes of ethics are developed for emerging work in machine learning and artificial intelligence, it may be useful to look toward those fields, like social work, that have long-standing histories and specific language for navigating social inequality. In a blog post, Catherine adapted the National Association of Social Workers Code of Ethics and replaced *social worker* with *data scientist* as a way of speculating about whether design and

forms of human understanding across power differentials, rather than using computation to replace human relationships. This kind of social innovation often goes underappreciated in the unicorn-wizard-genius model of data science. (For more on unicorns, see chapter 5.) As Patton says, "We had a lot of challenges with publishing papers in data science communities about this work, because it is very clear to me that they're slow to care about context. Not that they don't care, but they don't see the innovation or the social justice impact that the work can have."[49] Hopefully that will change in the future, as the work of SAFElab and others demonstrates the tremendous potential of combining social work and data science.

Communication Context

It's not just in the stages of data acquisition or data analysis that context matters. Context also comes into play in the

---

technical fields might ever be able to deal so explicitly with concepts of justice and oppression. Catherine D'Ignazio, "How Might Ethical Data Principles Borrow from Social Work?," Medium, September 2, 2018, https://medium.com/@kanarinka/how-might-ethical-data-principles-borrow-from-social-work-3162f08f0353.

49. Patton, interview by D'Ignazio.

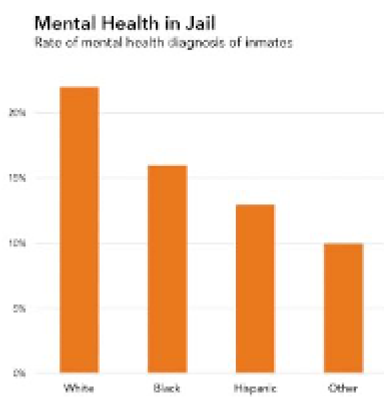
framing and communication of results. Let's imagine a scenario. In this case, you are a data journalist, and your editor has assigned you to create a graphic and short story about a recent research study: "Disparities in Mental Health Referral and Diagnosis in the New York City Jail Mental Health Service."[50] This study looks at the medical records of more than forty-five thousand first-time incarcerated people and finds that some groups are more likely to receive treatment, while others are more likely to receive punishment. More specifically, white people are more likely to receive a mental health diagnosis, while Black and Latinx people are more likely to be placed in solitary confinement. The researchers attribute some of this divergence to the differing diagnosis rates experienced by these groups before becoming incarcerated, but they also attribute some of the divergence to discrimination within the jail system. Either way, the racial and ethnic disparities are a product of structural racism.

Consider the difference between the two graphics shown in figure 6.6. The only variation is the title and framing of the chart.

Which one of these graphics would you create? Which one

50. Fatos Kaba, Angela Solimo, Jasmine Graves, Sarah Glowa-Kollisch, Allison Vise, Ross Macdonald, Anthony Waters, et al., "Disparities in Mental Health Referral and Diagnosis in the New York City Jail Mental Health Service," *American Journal of Public Health* 105, no. 9 (September 2015): 1911–1916.

should you create? The first—Mental Health in Jail—represents the typical way that the results of a data analysis are communicated. The title *appears* to be neutral and free of bias. This is a graphic about rates of mental illness diagnosis of incarcerated people broken down by race and ethnicity. The people are referred to as *inmates*, the language that the study used. The title does not mention race or ethnicity, or racism or health inequities, nor does the title point to what the data mean. But this is where additional questions about context come in. Are you representing only the four numbers that we see in the chart? Or are you representing the context from which they emerged?



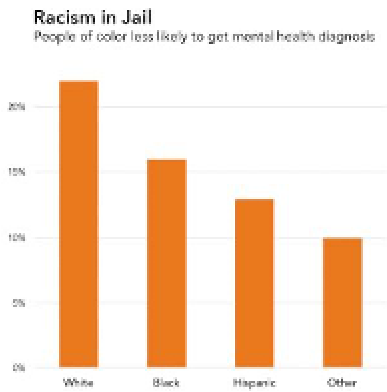


Figure 6: Two portrayals of the same data analysis. The data are from a study of people incarcerated for the first time in NYC jails between 2011 and 2013. Graphics by Catherine D'Ignazio. Data from Fatos Kaba et al., "Disparities in Mental Health Referral and Diagnosis in the New York City Jail Mental Health Service.

The study that produced these numbers contains convincing evidence that we should distrust diagnosis numbers due to racial and ethnic discrimination. The first chart does not simply fail to communicate that but also actively undermines that main finding of the research. Moreover, the language used to refer to people in jail as *inmates* is dehumanizing, particularly in the context of the epidemic of mass incarceration in the United States.[51] So, consider the second chart: Racism in Jail: People of Color Less Likely to Get Mental Health Diagnosis. This title offers a frame for how to interpret the numbers along the lines of the study from which they emerged. The research study was about racial disparities, so the title and content of this chart are about racial disparities. The people behind the numbers are *people*, not *inmates*. In addition, and crucially, the second chart names the forces of oppression that are at work: racism in prison.

Although naming racism may sound easy and obvious to some readers of this book, it is important to acknowledge that fields like journalism still adhere to conventions that resist such naming on the grounds that it is "bias" or "opinion." John Daniszewski, an editor at the Associated Press, epitomizes this view: "In general our policy is to try to be neutral and precise and as accurate as we possibly can be for the given situation. We're very cautious about throwing around accusations of our own that characterize something as being racist. We would try to say what was done, and allow the reader to make their own judgement."[52]

---

51. Prison reform advocate and formerly incarcerated person Eddie Ellis states that terms like *prisoner*, *inmate*, *convict*, and *felon* "are no longer acceptable for us and we are asking people to stop using them." See Eddie Ellis, "Language," Prison Studies Project, accessed July 29, 2019, http://prisonstudiesproject.org/language/.

52. Pete Vernon, "Dancing around the Word 'Racist' in Coverage of Trump,"

Daniszewski's statement may sound democratic ("power to the reader!"), but it's important to think about whose interests are served by making racism a matter of individual opinion. For many people, racism exists as a matter of fact, as we have discussed throughout this book. Its existence is supported by the overwhelming empirical evidence that documents instances of structural racism, including wealth gaps, wage gaps, and school segregation, as well as health inequities, as we have also discussed. Naming these structural forces may be the most effective way to communicate broad context. Moreover, as the data journalist in this scenario, it is your responsibility to connect the research question to the results and to the audience's interpretation of the results. Letting the numbers speak for themselves is emphatically not more ethical or more democratic because it often leads to those numbers being misinterpreted or the results of the study being lost. Placing numbers in context and naming racism or sexism when it is present in those numbers should be a requirement—not only for feminist data communication, but for data communication full stop.

This counsel—to name racism, sexism, or other forces of oppression when they are clearly present in the numbers—particularly applies to designers and data scientists

---

Columbia Journalism Review, September 25, 2017, https://www.cjr.org/covering_trump/trump-racism.php.

from the dominant group with respect to the issue at hand. White people, including ourselves, the authors of this book, have a hard time naming and talking about racism. Men have a hard time naming and talking about sexism and patriarchy. Straight people have a hard time seeing and talking about homophobia and heteronormativity. If you are concerned with justice in data communication, or data science more generally, we suggest that you practice recognizing, naming, and talking about these structural forces of oppression.[53]

But our work as hypothetical anti-oppression visualization designers is not over yet. We might have named racism as a structural force in our visualization, but there are still two problems with the "good" visualization, and they hinge on the wording of the subtitle: People of Color Less Likely to Get Mental Health Diagnosis. The first problem is that this is starting to look like a deficit narrative, which we discuss in chapter 2—a narrative that reduces a social group to negative stereotypes and fails to portray them with creativity and agency. The second issue is that by naming racism and then talking about people of color in the title, the graphic reinforces

53. As a nonhypothetical example of this, see the recent interactive feature from the *New York Times*, "Extensive Data Shows Punishing Reach of Racism for Black Boys," which models much of this advice in both naming racism and reflecting the findings of the study that served as the basis for the report. See https://www.nytimes.com/interactive/2018/03/19/upshot/race-class-white-and-black-men.html.

the idea that race is an issue for people of color only. If we care about righting the balance of power, the choice of words matters as much as the data under analysis. In an op-ed about the language used to describe low-income communities, health journalist Kimberly Seals Allers affirms this point: "We almost always use a language of deficiency, calling them disadvantaged, under-resourced and under-everything else. … It ignores all the richness those communities and their young people possess: the wealth of resiliency, tenacity and grit that can turn into greatness if properly cultivated."[54]

So let's give it a third try, with the image in figure 6.7.

In this third version, we have retained the same title as the previous chart. But instead of focusing the subtitle on what minoritized groups lack, it focuses on the unfair advantages that are given to the dominant group. The subtitle now reads, White People Get More Mental Health Services. This avoids propagating a deficit narrative that reinforces negative associations and clichés. It also asserts that white people have a race, and that they derive an unfair advantage from that race in this case.[55] Finally, the title is proposing an interpretation of

54. Kimberly Seals Allers, "What Privileged Kids—and Parents—Can Learn from Low-Income Youth," *Washington Post*, March 2, 2018.

55. How might we focus less attention on minoritized groups' disadvantages and more attention on dominant groups' unearned privileges? For example, instead of focusing on the women that are "missing" from data science and AI, perhaps we

the numbers that is grounded in the context of the researchers' conclusions on health disparities.

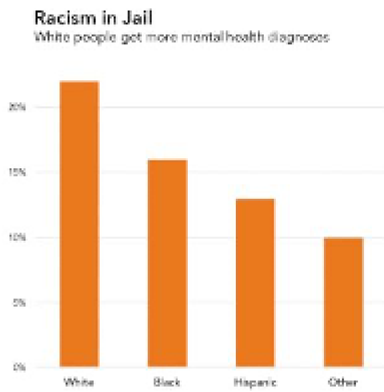


Figure 7: A third portrayal of the same data, with only the framing title and subtitle changed. Source: Data from Kaba et al., "Disparities in Mental Health Referral and Diagnosis in the New York City Jail Mental Health Service." Graphic by Catherine D'Ignazio. Data from Fatos Kaba et al., "Disparities in Mental Health.

---

should be focusing on the overabundance of men in data science and AI who don't see it as a problem worth their time and energy (because the system works for them).

Restoring Context

Three iterations on a single chart title might feel excessive, but it also helps to underscore the larger point that considering context always involves some combination of interest and time. Fortunately, there is a lot of energy around issues of context right now, and educators, journalists, librarians, computer scientists, and civic data publishers are starting to develop more robust tools and methods for keeping context attached to data so that it's easier to include in the end result.

For example, remember figure 6.3, that confusing chart of government procurements in São Paulo that we discussed earlier in this chapter? Gisele Craveiro, a professor at the University of São Paulo, has created a tool called Cuidando do Meu Bairro (Caring for My Neighborhood) to make that spending data more accessible to citizens by adding additional local context to the presentation of the information.[56] In the classroom, Heather Krause, a data scientist and educator, has developed the concept of the "data biography."[57] Prior to

56. See Gisele S. Craveiro and Andrés M. R. Martano, "Caring for My Neighborhood: A Platform for Public Oversight," in *Agent Technology for Intelligent Mobile Services and Smart Societies* (Berlin: Springer, 2014), 117–126.

57. See Heather Krause, "Data Biographies: Getting to Know Your Data," Global Investigative Journalism Network, March 27, 2017, https://gijn.org/2017/03/27/data-biographies-getting-to-know-your-data/.

beginning the analysis process, Krause asks people working with data, particularly journalists, to write a short history of a particular dataset and answer five basic questions: Where did it come from? Who collected it? When? How was it collected? Why was it collected? A related but slightly more technical proposal advocated by researchers at Microsoft is being called *datasheets for datasets*.[58] Inspired by the datasheets that accompany hardware components, computer scientist Timnit Gebru and colleagues advocate for data publishers to create short, three- to five-page documents that accompany datasets and outline how they were created and collected, what data might be missing, whether preprocessing was done, and how the dataset will be maintained, as well as a discussion of legal and ethical considerations such as whether the data collection process complies with privacy laws in the European Union.[59]

Another emerging practice that attempts to better situate

58. Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford, "Datasheets for Datasets," ArXiv.org, July 9, 2018.

59. Likewise, James Zou and Londa Schiebinger advocate for standardized metadata to accompany AI training datasets that spells out demographics, geographic limitations, and relevant definitions and collection practices. Zou and Schiebinger, "AI Can Be Sexist and Racist—It's Time to Make It Fair," *Nature*, July 18, 2018, https://www.nature.com/articles/d41586-018-05707-8.

data in context is the development of *data user guides*.[60] Bob Gradeck, manager of the Western Pennsylvania Regional Data Center, started writing data user guides because he got the same questions over and over again about popular datasets he was managing, like property data and 311 resident reports in Pittsburgh. Reports Gradeck, "It took us some time to learn tips and tricks. … I wanted to take the stuff that was in my head and put it out there with additional context, so other data users didn't have to do it from scratch."[61] Data user guides are simple, written documents that each contain a narrative portrait of a dataset. They describe, among other things, the purpose and application of the data; the history, format, and standards; the organizational context; other analyses and stories that have used the dataset; and the limitations and ethical implications of the dataset. This is similar to the work that data journalists are doing to compile datasets and then make them available for reuse. For example, the Associated Press makes comprehensive national statistics about school segregation in the United States available for purchase.[62] The

---

60. "Data User Guides," accessed August 6, 2019, http://www.wprdc.org/data-user-guides/.

61. Emerson Engagement Lab, "Civic Data Ambassadors: Module 2 Video 3—Civic Data Guides," video, 6:25, March 18, 2018, https://vimeo.com/260650894.

62. "School Segregation Data," ProPublica, December 2017,

spreadsheets are accompanied by a twenty-page narrative explainer about the data that includes limitations and sample story ideas.

These developments are exciting, but there is further to go with respect to issues of power and inequality that affect data collection environments. For example, professor of political science Valerie Hudson has worked for decades to trace the links between state security and the status of women. "I was interested in whether forms of oppression or subordination or violence against women were related to national, and perhaps international, instability and conflict," she explains. She and geographer Chad Emmett started the project WomanStats as a modest Excel spreadsheet in 2001. It has since grown to a large-scale web database with over a quarter of a million data points, including over 350 variables ranging from access to health care to the prevalence of rape to the division of domestic labor.[63]

Notably, their sources are qualitative as well as quantitative. Says Hudson, "If you want to do research on women, you have to embrace qualitative data. There's no two ways about it, because the reality of women's lives is simply not captured

---

https://www.propublica.org/datastore/dataset/school-segregation-charter-district-data.

63. WomanStats.org is solely focused on the status of women and does not collect any indicators on nonbinary people.

in quantitative statistics. Absolutely not."[64] At the present, WomanStats includes two types of qualitative variables: practice variables are composed from women's reports of their lived experiences, and law variables are coded from the legal frameworks in a particular country. Indeed, the WomanStats codebook is a context nerd's dream that outlines measurement issues and warns about the incompleteness of its own data, especially with respect to difficult topics.[65] In regard to the data that records reports of rape, for example—a topic upsetting enough to even consider, let alone contemplate its scale and scope in an entire country—the codebook states: "CAVEAT EMPTOR! Users are warned that this scale only reflects reported rape rates, and for many, if not most, countries, this is a completely unreliable indicator of the actual prevalence of rape within a society!"[66] Instead of focusing on a single variable, users are directed to WomanStats's composite scales, like the Comprehensive Rape Scale, which look at reported prevalence in the context of laws, whether laws are enforced, reports from lived experience, strength of taboos in that environment, and so on.

---

64. Valerie Hudson, interview with Catherine D'Ignazio, January 31, 2019.

65. "Codebook," WomanStats, accessed March 27, 2019, http://www.womanstats.org/new/codebook/.

66. "Codebook," http://www.womanstats.org/new/codebook/.

So tools and methods for providing context are being developed and piloted. And WomanStats models how context can also include an analysis of unequal social power. But if we zoom out of project-level experiments, what remains murky is this: Which actors in the data ecosystem are responsible for providing context?

Is it the end users? In the case of the missing Reddit comments, we see how even the most highly educated among us fail to verify the basic claims of their data source. And datasheets for datasets and data user guides are great, but can we expect individual people and small teams to conduct an in-depth background research project while on a deadline and with a limited budget? This places unreasonable expectations and responsibility on newcomers and is likely to lead to further high-profile cases of errors and ethical breaches.

So is it the data publishers? In the case of GDELT, we saw how data publishers, in their quest for research funding, overstated their capabilities and didn't document the limitations of their data. The Reddit comments were a little different: the dataset was provided by an individual acting in good faith, but he did not verify—and probably did not have the resources to verify—his claim to completeness. In the case of the campus sexual assault data, it's the universities who are responsible for self-reporting, and they are governed by their own bottom line.[67] The government is under-resourced to verify and document all the limitations of the data.

Is it the data intermediaries? Intermediaries, who have also been called *infomediaries*, might include librarians, journalists, nonprofits, educators, and other public information professionals.[68] There are strong traditions of data curation and management in library science, and librarians are often the human face of databases for citizens and residents. But as media scholar Shannon Mattern points out, librarians are often left out of conversations about smart cities and civic technology.[69] Examples of well-curated, verified and contextualized data from journalism, like the Associated Press database on school segregation or other datasets available in

67. Moreover, if one of the goals is transparency and accountability, the institutions in power often have strong incentives to *not* provide context, so the data setting is rife with conflicts of interest. Indeed, Gebru and colleagues foresee challenges to publishers specifying ethical considerations on their datasheets because they may perceive it as exposing themselves to legal and public relations risks. See Gebru et al., "Datasheets for Datasets."

68. Ricardo Ramírez, Balaji Parthasarathy, and Andrew Gordon, "From Infomediaries to Infomediation at Public Access Venues: Lessons from a 3-Country Study," in *Proceedings of the Sixth International Conference on Information and Communication Technologies and Development: Full Papers*, vol. 1 (New York: ACM, 2013), 124–132.

69. Shannon Mattern, "Public In/Formation," *Places Journal*, November 2016, https://placesjournal.org/article/public-information/.

ProPublica's data store, are also promising.[70] The nonprofit Measures for Justice provides comprehensive and contextualized data on criminal justice and incarceration rates in the United States.[71] Some data intermediaries, like Civic Switchboard in Pittsburgh, are building their own local data ecosystems as a way of working toward sustainability and resilience.[72] These intermediaries who clean and contextualize the data for public use have potential (and have fewer conflicts of interest), but sustained funding, significant capacity-building, and professional norms-setting would need to take place to do this at scale.

Houston, we have a public information problem. Until we invest as much in providing (and maintaining) context as we do in publishing data, we will end up with public information resources that are subpar at best and dangerous at worst. This ends up getting even more thorny as the sheer quantity of digital data complicates the verification, provenance, and contextualization work that archivists have traditionally undertaken. Context, and the informational infrastructure

---

70. "ProPublica Data Store," ProPublica, accessed August 6, 2019, https://www.propublica.org/datastore/.

71. See https://measuresforjustice.org/.

72. Aaron Brenner et al., "Civic Switchboard," accessed August 6, 2019, https://civic-switchboard.github.io/.

that it requires, should be a significant focus for open data advocates, philanthropic foundations, librarians, researchers, news organizations, and regulators in the future. Our data-driven lives depend on it.

Consider Context

The sixth principle of data feminism is to *consider context*. The bottom line for numbers is that they cannot speak for themselves. In fact, those of us who work with data must actively prevent numbers from speaking for themselves because when those numbers derive from a data setting influenced by differentials of power, or by misaligned collection incentives (read: pretty much all data settings), and especially when the numbers have to do with human beings or their behavior, then they run the risk not only of being arrogantly grandiose and empirically wrong, but also of doing real harm in their reinforcement of an unjust status quo.

The way through this predicament is by considering context, a process that includes understanding the provenance and environment from which the data was collected, as well as working hard to frame context in data communication (i.e., the numbers should not speak for themselves in charts any more than they should in spreadsheets). It also includes analyzing social power in relation to the data setting. Which power imbalances have led to silences in the dataset or data that is missing altogether? Who has conflicts of interest that prevent them from being fully transparent about their data? Whose knowledge about an issue has been subjugated, and how might we begin to recuperate it? The energy around context, metadata, and provenance is impressive, but until we fund context, then excellent contextual work will remain the exception rather than the norm.

WRAP UP

Key Takeaways

- Context is crucial in data science; ignoring it can lead to misleading interpretations and reinforce existing power imbalances.
- Ethical considerations in data science extend beyond the data itself to how it is framed and communicated, especially in terms of avoiding deficit narratives and being transparent about limitations.
- Data often comes from environments influenced by power differentials, and understanding this can help in identifying what is missing or misrepresented in a dataset.
- There are emerging tools and methods for

adding context to data, but these need to be more widely adopted and funded to become the norm rather than the exception.

Exercises

1. Analyze a dataset of your choice and write a "data biography" for it, answering questions like: Where did it come from? Who collected it? When? How was it collected? Why was it collected?

2. Find a data visualization online and critique it. Does it consider context? Does it avoid deficit narratives? Is it transparent about its limitations?

3. Create your own data visualization based on a dataset, making sure to provide context and to consider ethical implications. Write a brief

explanation of the choices you made in terms of framing and communication.

PART XIV
ALGORITHMS IN THE AGE OF CAPITALISM

Chapter Written by Robyn Caplan, Joan Donovan, Lauren Hanson, and Jeanna Matthews[1]

Learning Objectives

- Explain the concept of algorithmic

---

1. This chapter has been adapted from the Data & Society report Algorithmic Accountability: A Primer, originally by Robyn Caplan, Joan Donovan, Lauren Hanson, Jeanna Matthews. Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license. It has been modified to include learning objectives, key takeaways, and exercises.

accountability and its importance in societal decision-making.

- Understand the various ways in which algorithms can introduce or perpetuate bias, including through training data and design choices.
- Explain the challenges and complexities involved in auditing algorithms for fairness, transparency, and accountability.

WHAT IS AN ALGORITHM?

An algorithm is a set of instructions for how a computer should accomplish a particular task. Algorithms are used by many organizations to make decisions and allocate resources based on large datasets. Algorithms are most often compared to recipes, which take a specific set of ingredients and transform them through a series of explainable steps into a predictable output. Combining calculation, processing, and reasoning, algorithms can be exceptionally complex, encoding for thousands of variables across millions of data points. *Critically, there are few consumer or civil rights protections that limit the types of data used to build data profiles or that require the auditing of algorithmic decision-making*. Standards and enforcement for fairness, accountability, and transparency are long overdue for algorithms that allocate housing, healthcare, hiring, banking, social services, as well as goods and service delivery (Eubanks, 2018). Algorithmic accountability is the process of assigning responsibility for harm when algorithmic decision-making results in discriminatory and inequitable outcomes.

How Are Algorithms Used to Make Decisions?

Algorithmic decision-making is becoming more common every day. Increasingly, important decisions that affect people's lives are governed by datasets that are too big for an individual to process. People have become accustomed to algorithms making all manner of recommendations, from products to buy, to songs to listen to, to social network connections. But, algorithms are not just recommending, they are also being used to make big decisions about people's lives. Among many applications, algorithms are used to:

- Organize social media feeds;
- Display ads;
- Sort résumés for job applications;
- Allocate social services;
- Decide who sees advertisements for open positions,
- housing, and products;
- Decide who should be promoted or fired;
- Estimate a person's risk of committing crimes or the length of a prison term;
- Assess and allocate insurance and benefits;
- Obtain and determine credit; and
- Rank and curate news and information in search engines.

While algorithmic decision making can offer benefits in terms of speed, efficiency, and even fairness, there is a common misconception that algorithms automatically result in unbiased decisions. While it may appear like algorithms are unbiased calculations because they take in objective points of reference and provide a standard outcome, there remain many problems with those inputs and the outputs. As Frank Pasquale, law professor at the University of Maryland, points out, algorithmic decision-making is **"black boxed,"** which means that while we may know what goes into the computer for processing and what the outcome is, there are currently no external auditing systems or regulations for assessing what happens to the data during processing (Pasquale, 2015).

Algorithms are attractive because they promise neutrality in decision making—they take in data and deliver results. But algorithms are not "neutral." In the words of mathematician Cathy O'Neil, an algorithm is **an "opinion embedded in mathematics,"** (O'Neil, 2016). And like opinions, all algorithms are different. Some algorithms privilege a certain group of people over another. O'Neil argues that across a range of occupations, human decision makers are being encouraged to defer to software systems even when there is evidence that a system is making incorrect, unjust, or harmful decisions.

When an algorithm's output results in unfairness, we refer to it as bias. Bias can find its way into an algorithm in many ways. It can be created through the social context where an algorithm is created, as a result of technical constraints, or by the way the algorithm is used in practice (Friedman and Nissenbaum, 1996). When an algorithm is being created, it is structured by the values of its designer, which might not be neutral. And after an algorithm is created, it must be trained—fed large amounts of data on past decisions—to teach it how to make future decisions. If that training data is itself biased, the algorithm can inherit that bias. For these reasons and others, decisions made by computers are not fundamentally more logical and unbiased than decisions made by people.

Black-boxed algorithms can unfairly limit opportunities, restrict services, and even produce **"technological redlining."** As Safiya Noble, professor of communication at University of Southern California, writes, technological redlining occurs when algorithms produce inequitable outcomes and replicate known inequalities, leading to the systematic exclusion of Blacks, Latinos, and Native Americans (Noble, 2018). Technological redlining occurs because we have no control over how data is used to profile us. If bias exists in the data, it is replicated in the outcome. Without enforceable mechanisms of transparency, auditing, and accountability, little can be known about how algorithmic decision-making limits or impedes civil rights. Noble writes,

> technological redlining is a form of digital data discrimination, which uses our digital identities and activities to bolster inequality and oppression. It is often enacted without our knowledge, through our digital

> engagements, which become part of algorithmic, automated, and artificially intelligent sorting mechanisms that can either target or exclude us. It is a fundamental dimension of generating, sustaining, or deepening racial, ethnic, and gender discrimination, and it is centrally tied to the distribution of goods and services in society, like education, housing, and other human and civil rights. Technological redlining is closely tied to longstanding practices of 'redlining,' which have been consistently defined as illegal by the United States Congress, but which are increasingly elusive because of their digital deployments through online, internet-based software and platforms, including exclusion from, and control over, individual participation and representation in digital systems.[1]

Important examples of technological redlining were uncovered by ProPublica, who showed how Facebook's targeted advertising system allowed for discrimination by race and age (Angwin and Tobin, 2017; Angwin et al., 2017). These decisions embedded in design have significant ramifications for those who are already marginalized.

[1] Noble wrote this definition of "technological redlining" specifically for this publication.

Example: Racial Bias in Algorithms of Incarceration

One of the most important examples of algorithmic bias comes from the justice system, where a newly-created algorithmic system has imposed stricter jail sentences on black defendants. For decades, the company Northpointe has developed algorithmic systems for justice system recommendations. One such system is the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), which is used across the country to assess the risk of recidivism for defendants in pretrial hearings. The system operates on numerous points of data, such as questions about whether parents had separated and how many friends had been arrested, to make sentencing recommendations to judges. The goal of the system is to help balance protecting public safety while also eliminating the possible bias of human judges (Christin et al., 2015).

While the exact details of how COMPAS computes scores is proprietary information, the system has been built and tested across several dimensions by Northpointe's own team of computer scientists (Brennan et al., 2007; Brennan et al., 2009) and externally validated by researchers at Florida State University (Blomberg et al., 2010). Their analysis consistently showed that the system met a very commonly accepted definition of fairness within the field of statistics: (Chouldechova, 2016) for defendants of different races, it correctly predicted recidivism at about the same rate (Brennan et al., 2009; Blomberg et al., n.d.).

In 2016, however, ProPublica, a nonprofit news organization known for its investigative journalism, ran an analysis on how the system was being used in Broward County, Florida (Angwin et al., 2016). Their analysis revealed that even though the system predicted recidivism equally well for white and black defendants, it made different kinds of systematic mistakes for the two populations. **The system was more likely to mistakenly predict that black defendants were high-risk, while making the opposite type of mistake for white defendants.**

This meant that black defendants who would never go on to recidivate were being treated more harshly by the law, while white defendants who would go on to commit more crimes were being treated more leniently. To ProPublica, this was clear evidence of algorithmic bias (Angwin, 2016). Northpointe's response was to reassert the statistical merit of the COMPAS system. In the end, there were no public announcements made about changes to the COMPAS system, and it continues to be widely used within courts. The COMPAS conflict hinges on two key factors: there are no standard definitions for algorithmic bias, and there is no mechanism for holding stakeholders accountable.

Northpointe and ProPublica both agreed that COMPAS should meet some definition of racial fairness but neither agreed about what that meant. Because there was no public standard, Northpointe was free to create its own definition of fairness. When a challenge was made, Northpointe was not accountable to any particular set of values. Because of this lack of governance around the technologies of algorithmic risk assessment tools, the courts that continue to use the COMPAS system are not accountable either. Recently, the New York City Council passed a bill to determine a process for auditing the selection, use, and implementation of algorithms used by the city that directly affect people's lives ("The New York City Council – File #: Int 1696-2017", 2017). The bill highlights a need for assessment of disproportionate impacts across protected categories as well as a procedure for redress if harms are found.

COMPLICATIONS WITH ALGORITHMIC SYSTEMS

The COMPAS controversy demonstrates just how many different factors can complicate the design, use, assessment, and governance of algorithmic systems. Algorithms can be incredibly complicated and can create surprising new forms of risk, bias, and harm (Venkatsburamanian, 2015). Here, we lay out how complications in assessing fairness and bias are a result of stakeholders keeping algorithms intentionally opaque amidst calls for transparency. There is a need for greater reflection on models of power and control, where the sublimation of human decision-making to algorithms erodes trust in experts. Ultimately, regulators and researchers are ill-equipped to audit algorithms or enforce any regulation under these conditions.

Fairness and Bias

Algorithms are often deployed with the goal of correcting a source of bias in decisions made by humans. However, many algorithmic systems either codify existing sources of bias or introduce new ones. Additionally, bias can exist in multiple places within one algorithm.

An algorithmic system can take on unintended values that compete with designed values (Batya et al., 2006). In the case of COMPAS, the algorithm delivered discriminatory results because of the bias embedded in the training data. Because black people have historically been arrested at a higher rate than white people, COMPAS learned to predict that a black person is more at risk of being re-arrested than a white person. When implemented, this system reflects this learning back into the criminal justice system at a large scale, injecting a source of racial bias into steps of the judicial process that come after arrest.

By transferring values from one particular political and cultural moment to a different context, algorithms create a certain moral rigidity. Unless algorithms are consistently monitored and adjusted as time passes, they reinforce the values they were created with and can become rapidly outdated. For example, in terms of apportionment of healthcare, service delivery by insurance companies and hospitals depends on algorithmic decision-making, yet some doctors and caregivers do not agree with the standardized treatment models because these data are not robust enough to assess variables unavailable to the computer model, such as the unsteady living conditions of those in poverty.

Opacity and Transparency

Many algorithms are unable to be scrutinized because the data, process, or outcomes they rely on are kept behind closed doors. According to Jenna Burrell, this can happen for three reasons:

- Intentional corporate or state secrecy, such as a trade secrets;
- Inadequate education on the part of auditors; or
- Overwhelming complexity and scale on the part of the algorithmic system.

The more complex and sophisticated an algorithm is, the harder it is to explain, even by a knowledgeable algorithmic engineer.

Without some level of transparency, it is difficult to know whether an algorithm does what it says it does, whether it is fair, or whether its outcomes are reliable. For example, there is a clear-cut need for transparency around risk assessment tools like COMPAS, but this need is challenged by upholding trade secrets laws. Also, in some cases, transparency may lead to groups and individuals "gaming the system." For example, even the minimal openness surrounding how the trending feature on Twitter surfaces topics has allowed it to be manipulated into covering certain topics by bots and coordinated groups of individuals. Therefore, different contexts may call for different levels of transparency.

Repurposing Data and Repurposing Algorithms

Algorithms are expensive and difficult to build from scratch. Hiring computer scientists, finding training data, specifying the algorithm's features, testing, refining, and deploying a custom algorithm all cost time and money. Therefore, there is a temptation to take an algorithm that already exists and either modify it or use it for something it wasn't designed to do. However, accountability and ethics are context specific. Standards that were set and ethical issues that were dealt with in an algorithm's original context may be problems in a new application.

PredPol, a predictive policing service, uses an algorithm designed to predict earthquakes to find and assign police to hotspots (Huet, 2015). Crime data isn't the same as earthquake data, though, and civil rights organizations have criticized PredPol for using biased data to overpolice certain areas (Lartey, 2016). For a variety of reasons, crime data, especially that for arrests, is racially biased, which has an impact on any algorithm that uses it as training data.

This type of approach is also performed at an interpretive level, where the same data is interpreted to apply to a different context. For instance, credit history reports, which are designed to be evidence of financial responsibility, are often used as an input in hiring decisions, even though connections between credit history and work capability are dubious at best. In order to deal with such algorithmic creep, we may need new, more cost-effective systems for creating algorithms or more standards in place for evaluating when an algorithm can be successfully adapted from one application to another.

Lack of Standards for Auditing

Since the 1970s in the financial sphere, independent auditing has been used to detect instances of discrimination. While independent auditing could be used to detect bias in algorithmic systems, so far independent auditing is underutilized because of a lack of industry standards or guidelines for assessing social impact. One set of standards proposed by the Association for Computing Machinery US Public Policy Council seeks to ensure that automated decision-making is held to the same standards as equivalent human decision-making ("Statement on Algorithmic Transparency and Accountability," 2017). According to the ACM, these principles should be applied by algorithm designers at every stage in the creation process, putting the primary responsibility for their adoption in the hands of industry. Another set of guidelines, put forward by a coalition of industry and university researchers, advocates for social impact statements to accompany the sale and deployment of algorithmic products (Fairness, Accountability, and Transparency in Machine Learning, n.d.).

In the wake of the Facebook hearings, Russian disinformation campaigns, and the targeted harassment of civil rights organizers, civil society organizations, such as Color of Change and Muslim Advocates, are calling for independent audits of platforms and internet companies (Simpson, 2018). Data for Black lives has called for a "data public trust," where they ask Facebook to share anonymized datasets for public good (Milner, 2018). Data for Black Lives are also drafting a data code of ethics that would focus on data protections and limit digital profiling. Facebook reacted to Cambridge Analytica by deleting pages and limiting access to data, which forecloses the possibility of outside review (Facebook, 2018). As a result, it is imperative to create an organizational structure for independent auditing that is open and accessible to researchers and organizations.

Power and Control

One of the primary decisions made by algorithms is that of relevance of each dataset to other data points. What values, categories, and pieces of information are relevant to customers? Companies? States? Tarleton Gillespie (2014), a professor at Cornell University and principal researcher at Microsoft, states that algorithms are treated as trusted, objective sources of information. However, their decisions about relevance are choices shaped by a political agenda, whether that agenda is implicit or explicit to even the algorithm's own designers. This is especially important for algorithms that perform a gatekeeping role. Algorithms replicate social values but also embed them into systems, creating new standards and expectations for what is important in a given context. While there are laws prohibiting the sharing or sale of health and financial data by hospitals and banks, discrimination occurs because there are few protections in place for consumer data brokering, where discrete data points act as proxies for protected categories that are then assembled into profiles that are sold. This can lead to technological redlining.

Trust and Expertise

Trust means many things in different disciplines, but one sociological perspective holds that *trust is the belief that the necessary conditions for success are in place*. Those who are pro-algorithm suggests that humans are too trusting of other humans and some algorithms can outperform experts. Humans are accepting of error in other humans, but hold algorithms to a higher standard. In a series of studies conducted at the University of Chicago, researchers found that a subject's likelihood to use output from an algorithm dropped significantly after they saw evidence that the algorithm can make errors, even if it was still more accurate than their own responses. From this point of view, humans' lack of trust in algorithms is irrational. However, as Eubanks's and Noble's research shows, algorithms are just as capable of bias as humans, as they are embedded with subjective values.

Who is being endowed with trust has a direct relationship with where liability for decision making should fall. One way of avoiding responsibility is to keep an air of mystery around who is ultimately accountable through a lack of specification. In the COMPAS case, it wasn't clear who was liable for decisions so no one was held accountable for bias in the system. However, this can lead to a "moral crumple zone," where one entity is held legally liable for errors, even if they aren't in full control of the system (Elish and Hwang, 2015). For example, airplane pilots are held liable for the behavior of planes, even though many of the decisions are regularly made by computerized systems. Determining who is the trusted decision-maker between algorithmic engineers, algorithms, and users requires careful consideration of what the algorithm claims to do and who suffers from the consequences of mistakes. When an algorithm is making decisions or helping an expert make decisions, it becomes unclear who is ultimately responsible for the effects of those decisions.

WHAT IS ALGORITHMIC ACCOUNTABILITY?

Algorithmic accountability ultimately refers to the assignment of responsibility for how an algorithm is created and its impact on society; if harm occurs, accountable systems include a mechanism for redress. Algorithms are products that involve human and machine learning. While algorithms stand in for calculations and processing that no human could do on their own, ultimately humans are the arbiters of the inputs, design of the system, and outcomes. Importantly, the final decisions to put an algorithmic system on the market belongs to the technology's designers and company.

Critically, algorithms do not make mistakes, humans do. Especially in cases of technological redlining, assigning responsibility is critical for quickly remediating discrimination and assuring the public that proper oversight is in place. In addition to clearly assigning responsibility for the implementation of decisions made by algorithms, accountability must be grounded in enforceable policies that begin with auditing in pre- and post- marketing trials as well as standardized assessments for any potential harms. Currently, it is difficult to get technology corporations to answer for the harms their products have caused.

Below we outline how journalists, in consultation with academics and whistleblowers, have taken up the role of auditing algorithms, while also showing how *the lack of enforceable regulation led to a deficit in consumer protections.*

Auditing by Journalists

Currently, journalists are an important watchdog for algorithmic bias. Data journalism blends investigative methods from journalism with technical know-how to provide clear and accurate reporting on computational topics. While many algorithms are proprietary information, skilled journalists can use techniques of "reverse-engineering" to probe what's inside the black box by pairing inputs with outputs. A second approach facilitated by journalists is that of collaborative research with academics and whistleblowers. Particularly for personalization algorithms, which can be difficult or impossible to parse from the perspective of an individual user's account, peer-sourced research can reveal patterns that give clues about how the underlying algorithms work.

Enforcement and Regulation

The governance of algorithms is played out on an ad hoc basis across sectors. In some cases, existing regulations are reinterpreted to apply to technological systems and guide behavior, as with Section 230 of the Communications Decency Act. These instances can be hotly contested as algorithmic systems bring up new issues not before properly covered by the logic of existing precedents. In other cases, specific governing bodies are convened in order to set standards. For example, the Internet Governance Forum has been convened annually by the United Nations since 2006 and attempts to set non-binding guidance around such facets of the internet as the diversity of media content.

However, for accountability to be meaningful, it needs to come with the appropriate governance structures. According to Florian Saurwein, Natascha Just, and Michael Latzer, governance is necessary because algorithms impose certain risks, such as the violation of privacy rights and social discrimination (Saurwein et al., 2015). These risks need to be dealt with by the appropriate governance structure, which currently involves little oversight by states. Governance can occur by market and design solutions, such as product innovation that mitigates risk or consumers' ability to substitute risky products for ones they deem safer. Governance can also come from industry self-regulation, where company principles and collective decision-making favor public interest concerns. Last is traditional state intervention through mechanisms such as taxes and subsidies for certain kinds of algorithmic behavior. The appropriate structure must be matched with the context at hand to ensure the accountability mechanisms are effective.

Because of the ad hoc nature of self-governance by corporations, few protections are in place for those most affected by algorithmic decision-making. Much of the processes for obtaining data, aggregating it, making it into digital profiles, and applying it to individuals are corporate trade secrets. This means they are out of the control of citizens and regulators. As a result, there is no agency or body currently in place that develops standards, audits, or enforces necessary policies.

While law has always lagged behind technology, in this instance technology has become de facto law affecting the lives of millions—a context that demands lawmakers create policies for algorithmic accountability to ensure these powerful tools serve the public good.

WRAP UP

Key Takeaways

- Algorithms are not neutral; they can encode biases present in their training data or in the values of their designers, affecting decisions in areas like criminal justice, healthcare, and employment.
- The lack of standardized definitions for algorithmic fairness and the absence of regulatory oversight make it difficult to hold companies and organizations accountable for biased or harmful algorithms.
- Transparency in algorithmic decision-making is complicated by factors such as trade secrets, complexity, and the potential for system manipulation.

- Journalists, in collaboration with academics and whistleblowers, have become important watchdogs in auditing algorithms, but there is a need for formal governance structures.

Exercises

1. Discuss a real-world example of algorithmic bias and explore how it could be mitigated. What challenges would you anticipate in implementing these changes?
2. Conduct a mock audit of a hypothetical algorithm used for job recruitment. What criteria would you use to assess its fairness and accountability?
3. Debate the pros and cons of algorithmic decision-making in a specific sector (e.g., healthcare, criminal justice, or advertising).

How do you weigh the benefits of efficiency and scale against the risks of bias and lack of accountability.

REFERENCES

Angwin, Julia and Ariana Tobin. 2017. “Facebook (Still) Letting Housing Advertisers Exclude....” ProPublica. November 21, 2017. https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin

Julia Angwin et. al., “Machine Bias,” ProPublica, May 23, 2016, https://www.propublica.org/ article/machine-bias-risk-assessments-in-criminal-sentencing.

Angwin, Julia, Noam Scheiber, and Ariana Tobin. 2017. “Facebook Job Ads Raise Concerns About Age Discrimination.” The New York Times, December 20, 2017, sec. Business Day. https://www.nytimes.com/2017/12/20/business/ facebook-job-ads.html.

Batya Friedman, Peter H. Kahn Jr, and Alan Borning, “Value Sensitive Design and Information Systems,” in Human-Computer Interaction and Management Information Systems: Foundations, ed. Ping Zhang and Dennis F. Galletta (Abingdon: Routledge, 2006): 348-372.

Blomberg, Thomas, William Bales, Karen Mann, Ryan Meldrum, and Joe Nedelec. “Validation of the COMPAS Risk Assessment Classification Instrument.” College of Criminology and Criminal Justice, Florida State University, Tallahassee, FL, 2010. http://criminology.fsu.edu/ wp-content/uploads/Validation-of-the-COMPAS-Risk-Assessment-Classification-Instrument.pdf.

Brennan, Tim, Bill Dieterich, Beate Ehret, “Research Synthesis: Reliability and validity of COMPAS,” Northpointe Inc., September, 2007

Brennan, Tim, William Dieterich, and Beate Ehret. “Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System.” Criminal Justice and Behavior 36, no. 1 (January 2009): 21–40. https://doi.org/10.1177/0093854808326545

Chouldechova, Alexandra. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.” arXiv Preprint arXiv:1610.07524, 2016. https://arxiv.org/ abs/1610.07524

Christin, Angele, Alex Rosenblat, and danah boyd. “Courts and Predictive Algorithms.” CRIMINAL Justice Policy Program 38 (2015). http://www.datacivilrights.org/pubs/2015-1027/ Courts_and_Predictive_Algorithms.pdf.

Elish, Madeleine and Tim Hwang, “When Your Self-Driving Car Crashes, You Could Still be the One Who Gets Sued,” Quartz, July 25, 2015, https://qz.com/461905/when-your-self-drivingcar-crashes-you-could-still-be-the-one-who-gets-sued/

Eubanks, Virginia. 2018. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York, NY: St. Martin’s Press.

Facebook. "An Update on Our Plans to Restrict Data Access on Facebook | Facebook Newsroom." Published April 4 2018. https://newsroom.fb.com/news/2018/04/restricting-data-access/.

Fairness, Accountability, and Transparency in Machine Learning. n.d. "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms :: FAT ML." Accessed April 11, 2018. https://www.fatml.org/resources/principles-for-accountable-algorithms.

Friedman, Bayta and Helen Nissenbaum, "Bias in Computer Systems," ACM Transactions on Information Systems 14, no. 3 (1996): 330-347.

Gillespie, Tarleton. "The Relevance of Algorithms." Media Technologies: Essays on Communication, Materiality, and Society 167. (2014)

Huet, Ellen. 2015. "Server and Protect: Predictive Policing Firm PredPol Promises to Map Crime Before It Happens." Forbes. Accessed April 10, 2018. https://www.forbes.com/sites/ellenhuet/2015/02/11/predpol-predictive-policing/

Lartey, Jamiles. "Predictive Policing Practices Labeled as 'flawed' by Civil Rights Coalition." The Guardian, August 31, 2016. http://www.theguardian.com/us-news/2016/aug/31/predictive-policing-civil-rights-coalition-aclu.

Milner, Yeshimabeit, "An Open Letter to Facebook from the Data for Black Lives Movement." Medium (blog), April 4, 2018. https://medium.com/@YESHICAN/an-open-letter-to-facebook-from-the-data-for-black-lives-movement-81e693c6b46c.

Noble, Safiya Umoja. 2018. Algorithms of Oppression: How Search Engines Reinforce Racism. 1 edition. New York: NYU Press.

O'Neil, Cathy. 2016. Weapons of Math Destruction. Crown.

Pasquale, Frank. 2015. The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press.

Saurwein, Florian, Natascha Just, and Michael Latzer. "Governance of Algorithms: Options and Limitations." Info 17, no. 6 (September 14, 2015): 35–49. https://doi.org/10.1108/info05-2015-0025.

Simpson, Scott. "Muslim Advocates and Color Of Change Demand Independent Civil Rights Audit of Facebook." Muslim Advocates, April 3, 2018. https://www.muslimadvocates.org/ muslim-advocates-and-color-of-change-demand-independent-civil-rights-audit-of-facebook/.

"Statement on Algorithmic Transparency and Accountability," ACM US Public Policy Council, January 12, 2017, https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf.

"The New York City Council – File #: Int 1696-2017." 2017. http:// legistar.council.nyc.gov/LegislationDetail.aspx?ID=3137815&GUID=437A6A6D-62E1-47E2- 9C42-461253F9C6D0

Venkatsburamanian, Suresh . "When an algorithm isn't," Medium, October 1, 2015, https://medium.com/@geomblog/when-an-algorithm-isn-t-2b9fe01b9bb5

PART XV
RECOMMENDED READING (AND LISTENING/VIEWING)

Overview of Big Data

- Michael Keller and Josh Neufeld (2014), Terms of Service
- The Human Face of Big Data documentary
- Living in a Culture of Algorithms, Data & Society Podcast Episode 11
- Alex Pentland (2015), Social Physics: How Social Networks Can Make Us Smarter

Capitalism and Data

- Cathy O'Neil (2017), Weapons of Math Destruction
- Weapons of Math Destruction, Data & Society Podcast Episode 8

- Claudo Minca and Maartje Roelofsen (2021), "Becoming Airbnbeings: on datafication and the Quantified Self in Tourism" in Tourism Geographies
- Andrew McStay (2017), Privacy and the Media.
- "Nosedive," Black Mirror Season 3, Episode 1
- Adtech and the Attention Economy, Data & Society Podcast Episode 74
- Becoming Data Episode 5: Data & Racial Capitalism, Data & Society Podcast Episode 85
- Viktor Mayer-Schönberger and Thomas Range (2018), Reinventing Capitalism in the Age of Big Data
- Shoshana Zuboff (2019), The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power

Bias and Inequality

- Catherine D'Ignazio and Lauren F. Klein (2020), Data Feminism
- Data Feminism, Data & Society Podcast Episode 67
- Safiya Noble (2018), Algorithms of Oppression
- Algorithms of Oppression, Data & Society Podcast Episode 46
- Teri Schnaubelt (2018), Automating Inequality
- Coded Bias documentary
- Caroline Criado Perez (2019), Invisible Women: Data

Bias in a World Designed for Men

Artificial Intelligence

- Kai-Fu Lee (2018), AI Superpowers
- Jill Walker Rettberg (2022), “ChatGPT is Multilingual but monocultural, and it’s learning your values,” in jill/txt
- Kate Crawford (2021), Atlas of AI
- Person of Interest
- “Be Right Back,” Black Mirror, Season 2, episode 1
- Becoming Data Episode 3: Data, AI & Automation, Data & Society Podcast Episode 83
- Can ChatGPT Make This Podcast? Hard Fork Podcast
- AI Text Generators: Sources to Stimulate Discussion among Teachers
- Kai-Fu Lee and Chen Qiufan (2021), AI 2041: Ten Visions for Our Future

Politics and Computational Warfare

- Kai Strittmatter (2020), We Have Been Harmonized
- Samuel C. Woolley and Philip N. Howard (Eds.) (2019), Computational Propaganda

- Yochai Benkler Robert Faris, and Hal Roberts (2018), Network Propaganda
- The Great Hack
- A Look Inside China's Social Credit System
- Electionland Misinformation, Data & Society Podcast Episode 75
- Digital Technology and Democratic Theory, Data & Society Podcast Episode 78
- Peter Pomerantsev (2019), This Is Not Propaganda: Adventures in the War Against Reality
- Philip N. Howard (2020), Lie Machines: How to Save Democracy from Troll Armies, Deceitful Robots, Junk News Operations, and Political Operatives
- Jenny Goldstein and Eric Nost (Eds.) (2022), The Nature of Data: Infrastructure, Environments, Politics

My Related Writing

- Sylvia IV, J.J. and Kyle Moody (2019), "False Information Narratives: The IRA's 2016 Presidential Election Facebook Campaign." In Chiluwa, Innocent and Sergei Samoilenko (Eds.), Handbook of Research on Deception, Fake News and Misinformation Online, (pp. 326-348). IGI Global.
- Carrigan, Mark and J.J. Sylvia IV. 2022. "Is It Paranoia? A Critical Approach to Platform Literacy." *Journal of*

Media Literacy, The Human-Algorithmic Question: A Media Literacy Education Exploration Special Issue.

- Sylvia IV, J.J. (2016), "Little Brother: How Big Data Necessitates an Ethical Shift from Privacy to Power." In Booth, P. and A. Davisson, (Eds.), Controversies in Digital Ethics (pp. 13-28). Bloomsbury.
- Sylvia IV, J.J. (2010), "The Ethical Implications of A/B and Multivariate E-Commerce Optimization Testing". In Palmer, D.E. (Ed.), Ethical Issues in E-Business: Models and Frameworks (pp. 91-104). Business Science Reference.
- Sylvia IV, J.J. 2021. "An Affirmative Approach to Teaching Critical Data Studies." *Journalism and Media, Algorithms and Artificial Intelligence in Journalism and Media Special Issue.* Vol. 2, Issue 4. 641-656.
- Sylvia IV, J.J. (2020), "The Biopolitics of Social Distancing." in Social Media + Society, 2k Special Issue. Vol. 6, Issue 3.
- Sylvia IV, J.J. and Mark Andrejevic. (2016), "The Future of Critique: Mark Andrejevic on Power/Knowledge and the Big Data-Driven Decline of Symbolic Efficiency." International Journal of Communication. Vol. 10. 3230-3240.
- Sylvia IV, J.J. (Pre-Print) "From LiveJournal with Love: A Comparative Analysis of Russia's Domestic and International Disinformation Campaigns."
- Sylvia IV. J.J. (Draft) "Caught in the Middle:

LiveJournal's Geopolitical-Fueled Decline."

ORIGINAL CONTRIBUTORS

Aboubacar Camara is a sophomore at Fitchburg State University studying game design. He likes to talk and write about this field because it allows him to embrace his creativity and make something out of it. Aboubacar is very passionate about game design and likes to explore different aspects of the subject. He has worked on a few projects consisting of first-person shooters and platform levels. One of the first-person shooter levels was based on an office building and the other was based on an abandoned military base. From his peers, Aboubacar is said to be very creative in the things he works on. He also is known for being empathetic towards his peers. What Aboubacar wishes to do after graduation, is to work for a game development company or start a small one. He believes that with the skills he obtained from Fitchburg State, he will be able to become a game developer.

Ana'aya McGowan Mozell is a sophomore at Fitchburg State University, majoring in PR, Social Media & Advertising.

Brendan Smith is a junior at Fitchburg State University, where he studies Film and Video. He is an aspiring filmmaker who is interested in producing, directing, and assistant directing. In his free time, he researches and watches films from around the world in hopes of gaining both inspiration from them and giving them the proper recognition they deserve for the impact that they have had on the world of filmmaking.

Email: brensmith259@gmail.com

Elise Takehana teaches writing and digital storytelling in the English Studies department at Fitchburg State University. She also co-founded the Digital Media Innovation program at FSU and worked with colleagues to create digital archive exhibits from the Robert E. Cormier collection. Her own research focuses on the impact of the medium on storytelling and style.

Glenndale Bartolome is a sophomore at Fitchburg State university, studying Game Design. He has at this point developed a few levels, assisted in the creation of a few small games, and participated in a couple of Game Jams, taking minor roles but slowly trying to be more involved as he progresses through the school year. One game he got to test was a VR shooter called "Worm Punk", getting a beta view of the game's map. He hopes to be able to land a job at an independent game company that treats its workers fairly, and help enforce a love for video games. He generally prefers privacy above all else, not being comfortable with sharing things like contact information, or even calling strangers and unknown numbers on the phone. As such, his contact info will not be provided, save for his school email.

Email: gbartolo@student.fitchburgstate.edu

Henry Christiansen is a student at Fitchburg State University studying film and video production and currently in his sophomore year. Some of his achievements include a short film that he wrote, filmed, and produced. Middlesex Community College awarded him an honorable mention. He was also an intern at Middlesex Community College for film and video production. Some of his career goals after graduation are to be a cinematographer or a production manager. Since being in school, he has worked really hard on every project he has been a part of. He currently works full-time in a grocery store to help pay for college and other family needs. A fun fact about him is that he has worked with Keith Urban and Big Time Rush. He filmed them on stage and got to meet them which was an amazing experience.

Email: hchristiansen2018@gmail.com

LinkedIn: https://www.linkedin.com/in/henry-christiansen-4808811b5/

J.J. Sylvia IV (Ph.D., North Carolina State University) is an Associate Professor of Communications Media at Fitchburg State University, where he co-founded an undergraduate major in Digital Media Innovation and a master's program in Applied Communication, focusing on Social Media. The core of his research involves the philosophy of communication and the analysis of the impacts of big data, algorithms, and emerging media on processes of subjectivation — the ways we are shaped as subjects. Sylvia's academic training includes an M.A. in Philosophy and a Ph.D. in Communication, Rhetoric, and Digital Media. He lives with his wife and two daughters in Worcester, MA.

Contact: jsylvia3@fitchburgstate.edu

Keshauni Johnson is a Game Design major and a sophomore at Fitchburg State University. She aspires to open her own professional game development studio in the future. Some of her favorite hobbies include reading, drawing, and playing games. Her favorite video game of all time is Kirby's Return to Dreamland.

Leonora Shell (M.A.T., Mississippi University for Women) is a business owner, educator, entomologist, science communicator, and writer with a love of digital media and sourdough bread. She earned her B.S. in Entomology from the University of California at Davis. She has been the Curator of Digital Media with Your Wild Life and the Rob Dunn Lab at North Carolina State University, as well as the Digital Learning Specialist at the North Carolina Museum of Natural Sciences. She is currently the C.O.O. and co-owner of an e-commerce business.

Sophia Moore is a freshman studying communications media with a concentration in public relations and social media advertising at Fitchburg State University. She received Dean's list her first semester and has worked on several marketing projects throughout her first year of college. She aspires to have a career in sports marketing, motivated by her love for sports and interest in marketing. Growing up she was a multi-sports athlete and continues to play on the Fitchburg State women's soccer team.

Contact: s.j.moore744@gmail.com

GRANT INFORMATION

The U.S. Department of Education, the granting agency for the ROTEL (Remixing Open Textbooks through an Equity Lens) project, requires information about the grant be included in the back matter. The text for this section is provided below.

The contents of this book were developed under a grant from the Fund for the Improvement of Postsecondary Education, (FIPSE), U.S. Department of Education. However, those contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government.

For more information about the ROTEL Project, please visit our project website.

VERSION HISTORY

Below is the version history for The Data Renaissance: Analyzing the Disciplinary Effects of Big Data, Artificial Intelligence, and Beyond.

Version	Publication Date	Cha
First Edition: link to first edition	February 29, 2024	—
Revised Edition: link to revised edition	October 2, 2024	Adde Versi Chap