MACQUARIE University

An Ethical Framework for Data Science Ethics

Lecturer Dr. YANG ZHANG

MACQUARIE University

- 5 Ethical Principles
- Introduction to Data Science Ethics
- Real-world Examples
  -Ethical Errors in Infographic

Five Ethical Principles

MACQUARIE University

The Ethics of Data

MACQUARIE University

The ethics of data examines the moral challenges associated with gathering and analyzing extensive datasets, including concerns related to the application of big data.

https://www.kaggle.com/code/kmldas/data-ethics-in-data-science-analytics-ml-and-ai

Laws Related to the Protection of Data

MACQUARIE University

ANALYSIS: Algorithmic Bias Is No Longer Under Regulators' Radar (2)

Rachel DuFault
Legal Content Specialist

The Bloomberg Law 2023 series previews the themes and topics that our legal analysts will be watching closely in 2023. Our ESG & Employment analyses focus on two arenas where regulatory agendas and corporate practices are sure to clash in the year ahead.

Nationwide, 2023 is teed up to tackle the new frontier of tech bias in the workplace: algorithmic discrimination.

In 2022, the Equal Employment Opportunity Commission laid the foundation by rolling out guidance about workplace algorithmic bias. The issue is also becoming a priority for state and local lawmakers, as they draft and enforce new restrictions on discriminatory use of artificial intelligence in workplaces. And a model of industry self-regulation is in the works for AI and

https://news.bloomberglaw.com/bloomberg-law-analysis/analysis-algorithmic-bias-is-no-longer-under-regulators-radar

Laws Related to the Protection of Data

MACQUARIE University

Litigation | Attorney Analysis | Data Privacy

U.S. data privacy laws to enter new era in 2023

By **Fredric D. Bellamy**

January 13, 2023 2:21 AM GMT+11 · Updated a year ago

Commentary | Attorney Analysis from Westlaw Today, a part of Thomson Reuters.

https://www.reuters.com/legal/legalindustry/us-data-privacy-laws-enter-new-era-2023-2023-01-12/

Laws Related to the Protection of Data

General data protection regulation (GDPR)

The EU general data protection regulation (GDPR) was adopted in 2016 and entered into application in May 2018.

Thanks to the GDPR, there is one set of data protection rules for all companies operating in the EU, wherever they are based.

The stronger rules introduced by the GDPR mean that:

- **people have more control** over their personal data
- **businesses benefit** from a level playing field

Uniform and up-to-date legislation on data protection is essential to guarantee individuals' fundamental right to have their personal data protected, to allow for the development of the digital economy and to strengthen the fight against crime and terrorism.

› General data protection regulation (background information)

Infographic - Data protection regulation

See full infographic

https://www.consilium.europa.eu/en/policies/data-protection/

Laws Related to the Protection of Data

MACQUARIE University

- 1974, US Privacy Act - regulates *federal govt.* collection, use ,and disclosure of personal information.
- 1996, US Health Insurance Portability & Accountability Act (HIPAA) - protects personal health data.
- 1998, US Children's Online Privacy Protection Act (COPPA) - protects data privacy of children under 13.
- 2018, General Data Protection Regulation (GDPR) - provides user rights, data protection ,and privacy.
- 2018, California Consumer Privacy Act (CCPA) gives consumers more *rights* over their (personal) data.
- 2021, China's Personal Information Protection Law just passed, creating one of the strongest online data privacy regulations worldwide.

https://www.consilium.europa.eu/en/policies/data-protection/

Data Science Ethical Considerations

MACQUARIE University

American Statistical Association

Data Science Code of Professional Conduct

AI4 People's Framework

Ethics Challenges

MACQUARIE University

1. Data Ownership
2. Informed Consent
3. Intellectual Property
4. Data Privacy
5. Right To Be Forgotten
6. Dataset Bias
7. Data Quality
8. Algorithm Fairness
9. Misrepresentation
10. Free Choice

Case Study

Data Privacy

MACQUARIE University

Why 'Anonymous' Data Sometimes Isn't

Anonymous data sets are an enormous boon for researchers, but the recent de-anonymization of Netflix customer data shows there are privacy risks as well. Commentary by Bruce Schneier.

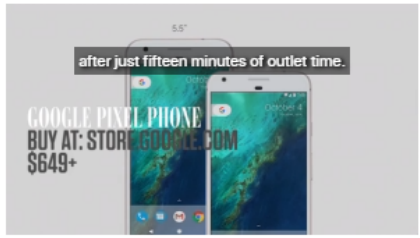
LAST YEAR, NETFLIX published 10 million movie rankings by 500,000 customers, as part of a challenge for people to come up with better recommendation systems than the one the company was using. The data was anonymized by removing personal details and replacing names with random numbers, to protect the privacy of the recommenders.

Arvind Narayanan and Vitaly Shmatikov, researchers at the University of Texas at Austin, de-anonymized some ofthe Netflix data by comparing rankings and timestamps with public information in the Internet Movie Database, or IMDb.

Their research (.pdf) illustrates some inherent security problems with anonymous data, but first it's important to explain what they did and did not do.

They did *not* reverse the anonymity of the entire Netflix dataset. What they did was reverse the anonymity of the Netflix dataset for those sampled users who also entered some movie rankings, under their own names, in the IMDb. (While IMDb's records are public, crawling the site to get them is against the IMDb's terms of service, so the researchers used a representative few to prove their algorithm.)

TRENDING NOW



Google Just Got Real By Changing Its Gadget Game

MOST POPULAR

CULTURE

https://www.wired.com/2007/12/why-anonymous-data-sometimes-isnt/

Case Study

Data Privacy & User Rights

MACQUARIE University

After Data Breach Exposes 530 Million, Facebook Says It Will Not Notify Users

APRIL 9, 2021 · 11:58 PM ET

Emma Bowman

The leaked data includes personal information from 533 million Facebook users in106 countries.
Olivier Douliery/AFP via Getty Images

Facebook decided not to notify over 530 million of its users whose personal data was lifted in a breach sometime before August 2019 and was recently made available in a public database. Facebook also has no plans to do so, a spokesperson said.

Phone numbers, full names, locations, some email addresses, and other details from user profiles were posted to an amateur hacking forum on Saturday, Business Insider reported last week.

https://www.npr.org/2021/04/09/986005820/after-data-breach-exposes-530-million-facebook-says-it-will-not-notify-users

Case Study

Collection Bias

MACQUARIE University

The Hidden Biases in Big Data

by Kate Crawford

April 01, 2013

This looks to be the year that we reach peak big data hype. From wildly popular big data conferences to columns in major newspapers, the business and science worlds are focused on how large datasets can give insight on previously intractable challenges. The hype becomes problematic when it leads to what I call "data fundamentalism," the notion that correlation always indicates causation, and that massive data sets and predictive analytics always reflect objective truth. Former *Wired* editor-in-chief Chris Anderson embraced this idea in his comment, "with enough data, the numbers speak for themselves." But can big data really deliver on that promise? Can numbers actually speak for themselves?

Sadly, they can't. Data and data sets are not objective; they are creations of human design. We give numbers their voice, draw inferences from them, and define their meaning through our interpretations. Hidden biases in both the collection and analysis stages present considerable risks, and are as important to the big-data equation as the numbers themselves.

https://hbr.org/2013/04/the-hidden-biases-in-big-data

Introduction to Infographics

MACQUARIE University

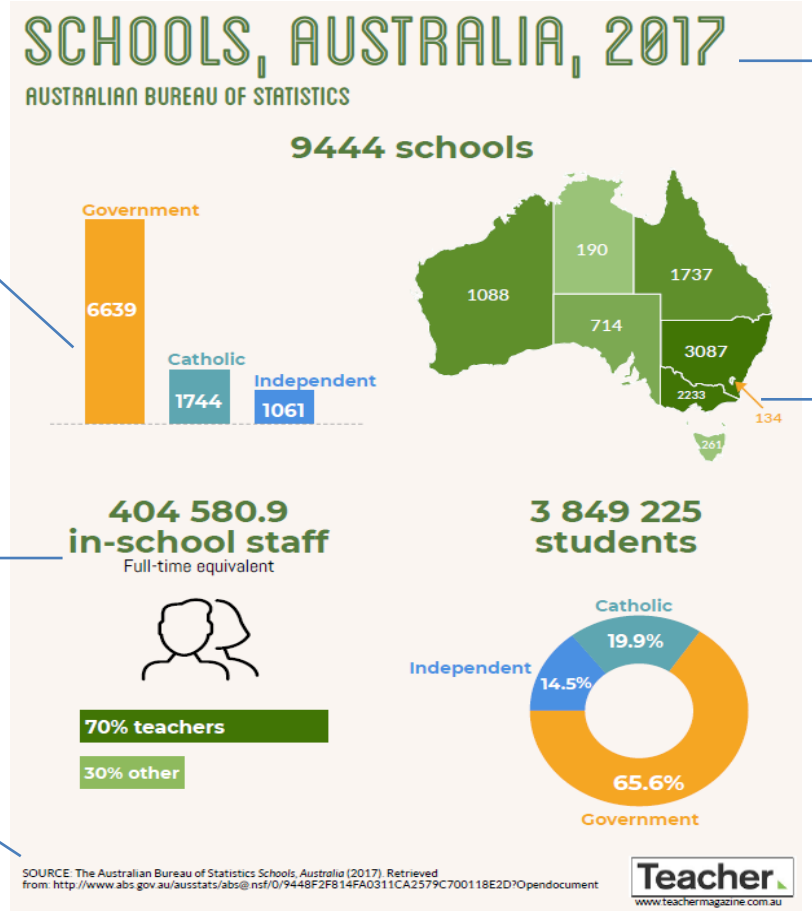
- An infographic is a visual communication tool to communicate complex data to the audience in an engaging way to easily interpret, understand and read the information.

- The infographic uses data visualisations building blocks such as illustrations, text, maps, icons, charts, and graphics to tell a cohesive story (Krum, 2013).

- The information in the infographics needs to have :
  - ***a clear purpose***
  - ***to be conveyed to an appropriate audience***
  - ***the message conveyed is not out of scope***

EXAMPLE OF AN INFOGRAPHIC

Infographics are visualization of data or ideas that tries to convey complex information to an audience in a manner that can be quickly consumed and easily understood.



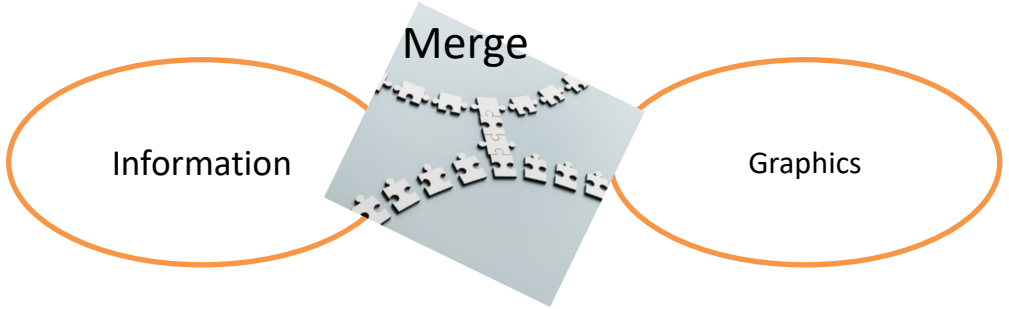
Data Representation

Heading

Statistics

Text

Source



(Toth and McClure, 2016)

Smiciklas, M. (2012). The power of infographics: Using pictures to communicate and connect with your audiences. Que Publishing.

Krum, R. (2013). Cool infographics: Effective communication with data visualization and design. John Wiley & Sons.

Rhetorical Dimensions of Infographics

MACQUARIE University

The producer needs to conduct a research process following a rhetorical strategy

- purpose
- topic
- determine the extent of information needed
- access source material
- evaluate the source material

(Toth and McClure, 2016)

Genre of Infographics

MACQUARIE University

"Graphics offers the producer a lot of rhetorical power". ~~(Toth and McClure, 2016, p. 264)

Responsibilities of the Producer

- accurate information
- credibility of the source → **ethical errors**
- avoid distortion
- use unbiased data

Ethical Errors in Infographic

MACQUARIE University

- There is a predisposition of individuals' trust when the infographics are from reputable sources.

- The underlying ethical errors in infographics can
  - ***impact the users' perceptions and interpretations***
  - ***impact the users' decision-making and judgement***
  - ***raise concerns about information credibility***
  - ***improperly communicate the indented message***
- Leading to ethical implications to humanity such as
  - ***harm, not beneficial, not fair, no choices, not responsible***

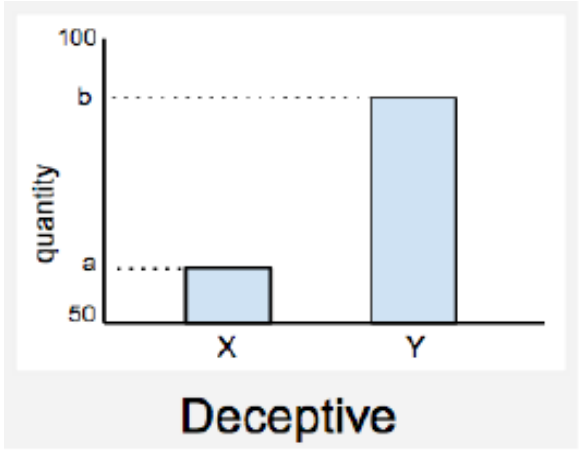
Deception in Infographics Leading to Ethical Errors

"GOOD DESIGN HAS TWO KEY ELEMENTS: SIMPLICITY OF DESIGN AND COMPLEXITY OF DATA."

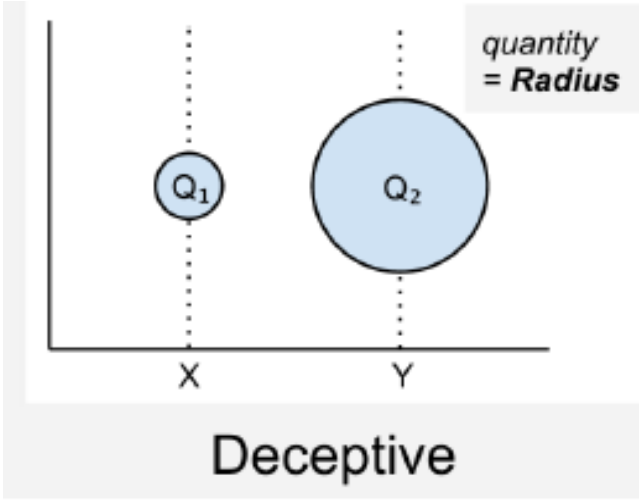
~~ (TUFTE P. 176)

Deceptions Levels:

- Chart Level



Deceptive



Deceptive

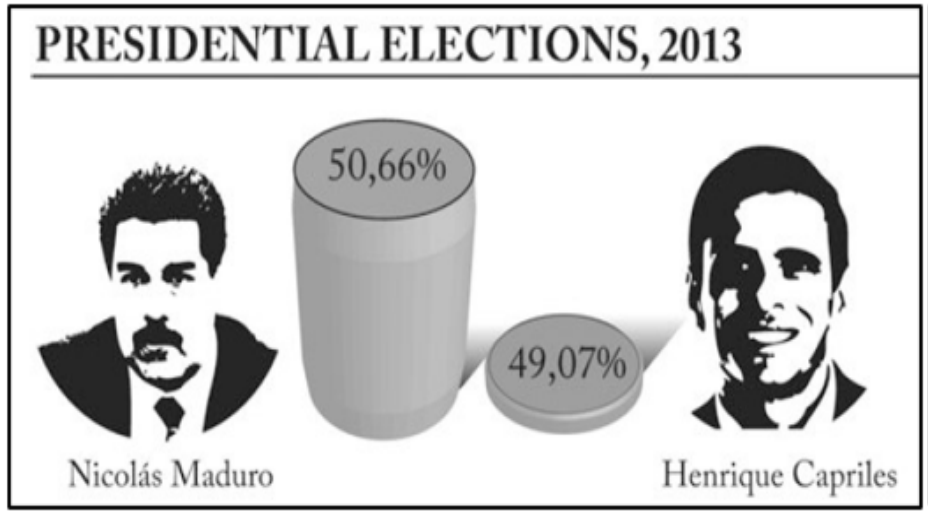
- Message Level

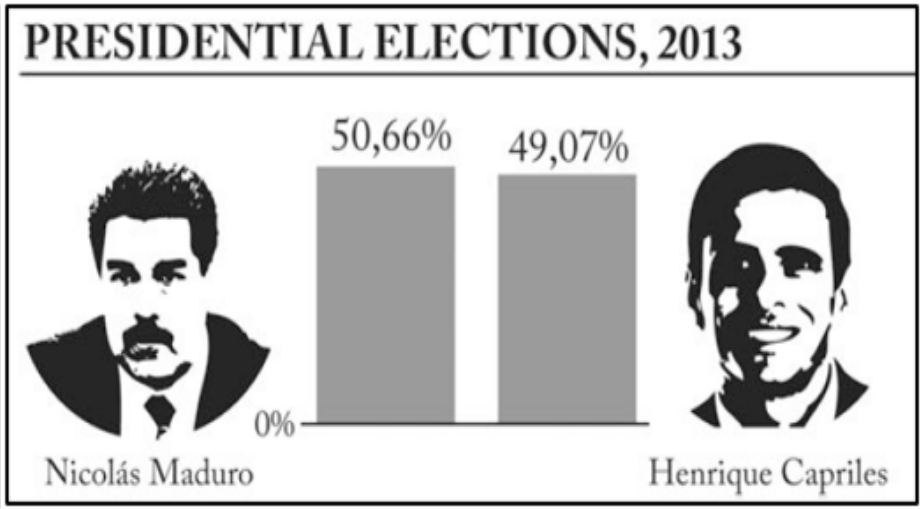
(Message reversal and message exaggeration/understatement)

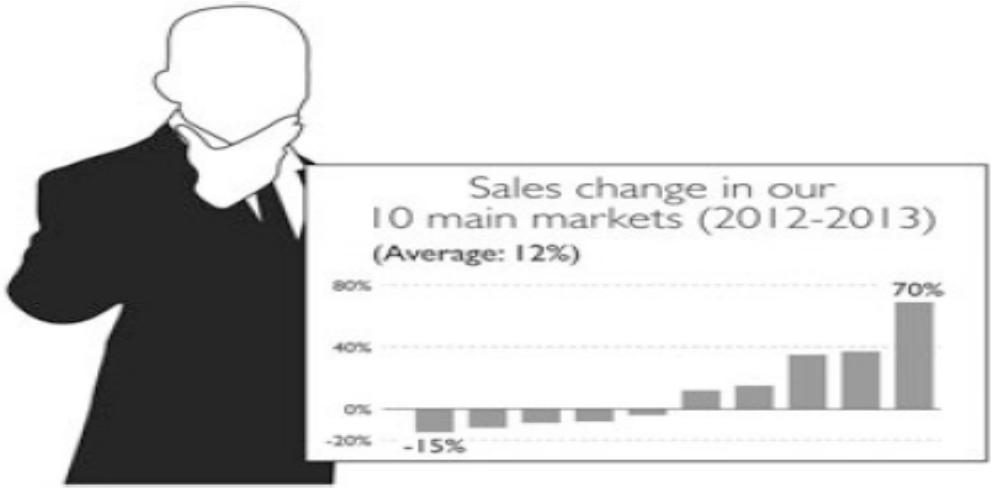
(Pandey et al., 2015)

Misleading Graphics

Examples: *







(Cairo, 2014a, 2014b) * originally from Venezuela de televisión 2013 presidential election results (figure on the left side)

Misleading Graphics –Cherry Picking

What is cherry picking?

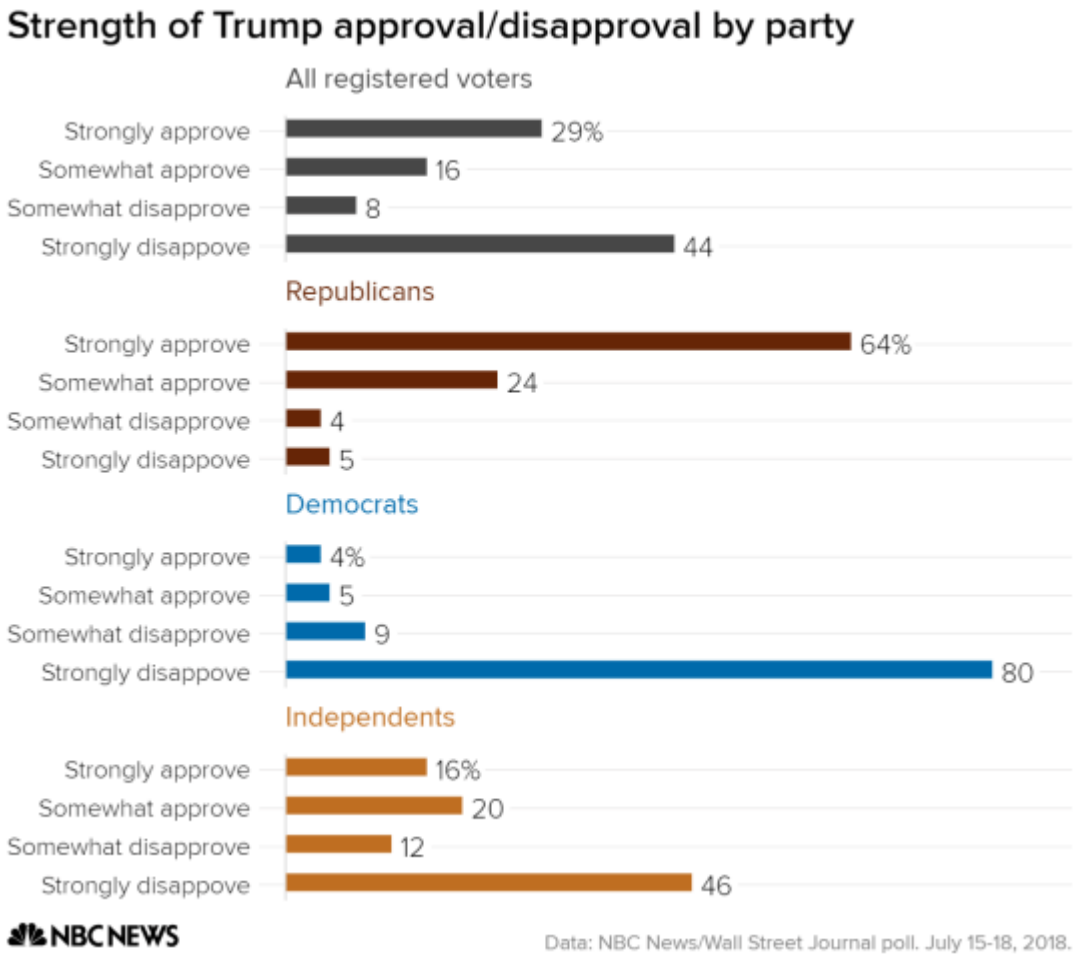
Cherry picking is the deliberate practice of presenting the results of a study or experiment that best support the hypothesis or argument, instead of reporting all the findings. Cherry picking can also be applied to the process of conducting an experiment, where the researcher intentionally uses limited categories of selection for their participants, in order to carry out their experiment.

Here is an example to better explain what cherry picking is; let's say a researcher does an experiment to prove a claim. After conducting the experiment, they find that only 20% of the results actually support their claim, while the other 80% disprove it. A researcher who decided to employ cherry picking would present only the 20% of their results that supported their claim, rather than all the results obtained in the experiment.

Cherry-picking is a form of data-driven deception where certain sets or sources or information are omitted from a survey, study, chart, or graph. The primary reason for cherry-picked visuals is an aim to offer clean, predictable results that fit into a neat trend, pattern, or box. The issue with cherry-picking is that it doesn't paint an honest, objective picture, offering results that are inaccurate or missing out on vital segments of knowledge.

http://ds-wordpress.haverford.edu/psych2015/projects/chapter/cherry-picking-data/

Misleading Graphics –Cherry Picking



Ethical Consideration

Ethical considerations across phases of data-driven visual story telling

- Data Acquisition
- Data Transformation
- Data Conveying and Connecting Insights

(Diakopoulos, 2018)

What To Do

Ethics Checklists in Data Science

MACQUARIE University

Here's a checklist for people who are working on data projects:

- ❑ Have we listed how this technology can be attacked or abused?
- ❑ Have we tested our training data to ensure it is fair and representative?
- ❑ Have we studied and understood possible sources of bias in our data?
- ❑ Does our team reflect diversity of opinions, backgrounds, and kinds of thought?
- ❑ What kind of user consent do we need to collect to use the data?
- ❑ Do we have a mechanism for gathering consent from users?
- ❑ Have we explained clearly what users are consenting to?
- ❑ Do we have a mechanism for redress if people are harmed by the results?
- ❑ Can we shut down this software in production if it is behaving badly?
- ❑ Have we tested for fairness with respect to different user groups?
- ❑ Have we tested for disparate error rates among different user groups?
- ❑ Do we test and monitor for model drift to ensure our software remains fair over time?
- ❑ Do we have a plan to protect and secure user data?

Use Ethics Checklists in Data Science Projects
Data scientists can use ethics checklists when conducting data science projects.
The checklist is a list of questions that helps determine whether a data scientist has considered various aspects of ethics.

https://resources.oreilly.com/examples/0636920203964/blob/master/of_oaths_and_checklists.md

MACQUARIE University

Thank you