SESUG Paper 079-2021

Developing Ethical Data Use and Users

Smith, Kelly D., Central Piedmont Community College

ABSTRACT

Just because we can, should we? Has the ability to analyze data outpaced the growth of data ethics? After a short review of ethics and the current state of data ethics, join a discussion of where we are, where we want to be, and how to get us there. Should data ethics training be required? What are the options to promote ethical data use and deter poor ethical choices?

INTRODUCTION

Ethics, the “moral principles that govern a person’s behavior” (Dictionary.com) has always interested me. How do people decide what to do, or perhaps more importantly, what *not* to do? In terms of data ethics, a discussion session at AIR Forum 2018 (Webber & Morn) sparked my interest with the title ”The Uses and Potential Misuses of Data.” That discussion focused mainly on ethical visualizations of data but data ethics covers much more than displaying data accurately and without distortion. Floridi and Taddeo (2016, p. 1) defined data ethics as a

> … new branch of ethics that studies and evaluates moral problems related to data (including generation, recording, curation, processing, dissemination, sharing and use), algorithms (including artificial intelligence, artificial agents, machine learning and robots) and corresponding practices (including responsible innovation, programming, hacking and professional codes), in order to formulate and support morally good solutions (e.g. right conducts or right values).

Data ethics is a broad field, reaching into every aspect of data science. In a recent essay, Schmarzo reminds data scientists that ethics can be passive (“do no harm”) or proactive (“do good”), and states that a proactive ethics approach is necessary because analytics models can do as much harm as much good (2020, p. 20).

DATA IS EVERYWHERE

The importance of data ethics grows in tandem with the increasing presence of data in every aspect of daily life. How much data is being produced? Best estimates suggest the amount of data produced each day is more than 2.5 exabytes (1 exabyte = 10^{18} bytes). Figure 1 is a portion of an infographic created by Price for CloudTweaks (n.d.) that attempts to connect large data amounts to daily life. The amount of data produced is growing exponentially; by 2025, new data is expected to reach 436 exabytes per day (Vuleta, 2021). Perhaps more importantly, the ease of collecting and storing of data is keeping pace with its generation. This means more and more data is available for exploration and analysis. The increasing sophistication and accessibility of data technology, including Artificial Intelligence and Machine Learning, add to the potential for data misuse.

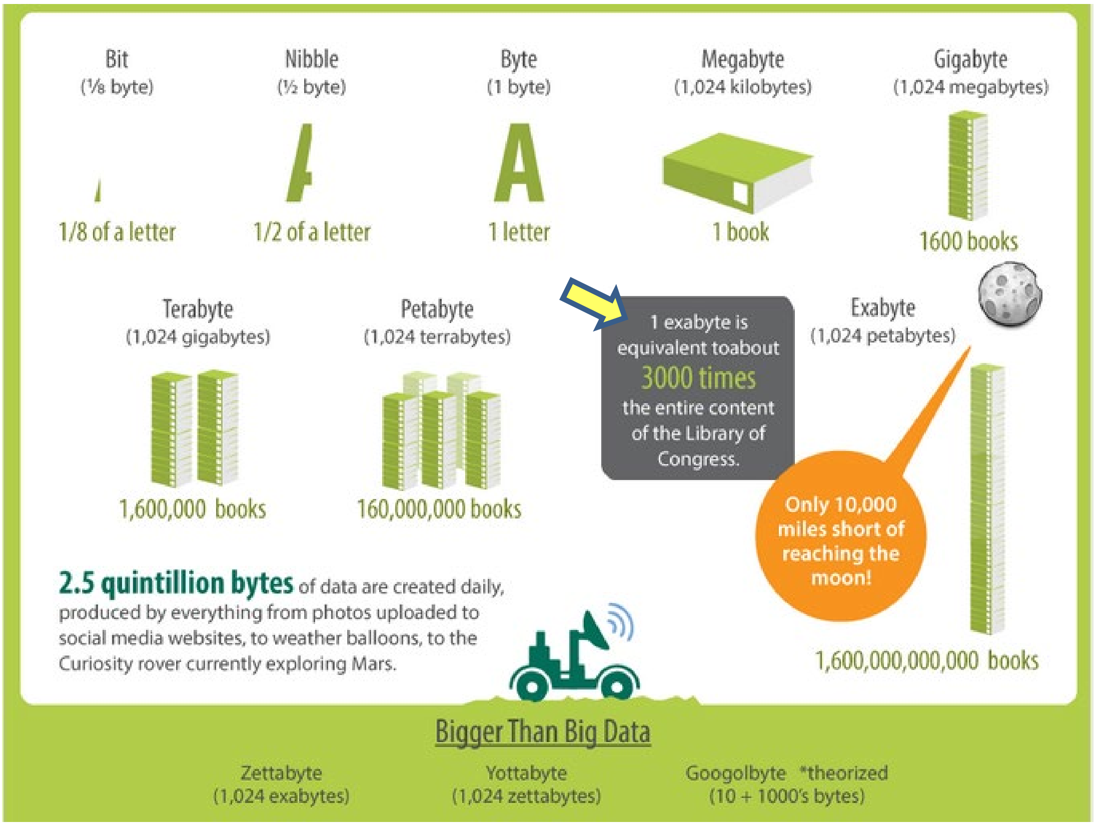

Figure 1. Connecting data generation to daily life. Portion of infographic from CloudTweaks. (n.d.) (arrow added to highlight exabyte definition)

DATA MISUSE

Hand (2018) pointed out "it is not the data *per se* that raise ethical issues, but the use to which they are put and the analysis to which they are subjected" (p. 177). In other words, data collected for one purpose may or may not be suitable for another purpose. Data misuse can occur when the questions analysts attempt to answer are not a good match with the data's original purpose. Two sayings that sum up the potential ways to misuse data are "Garbage In, Garbage Out" and "Just because we can doesn't mean we should."

INCIDENTAL MISUSE

Recently, Siegel described the incidental misuse of data: "We're not talking about mishandling, leaking, or stealing data. Rather, this is the generation of *new* data – the indirect discovery of *unvolunteered* truths about people [emphasis added]" (2020, para 1.) In 2012, the New York Times reported on Target's use of shopper data to predict if female customers were pregnant in order for Target to offer them coupons on pregnancy related items (Duhigg, 2012). The company was thinking about finances, not about the potential impact on its customers when deciding to use shopper purchasing data to create predictive models. Whether or not such an event actually occurred, Target's predictions could have caused family conflict when one member knew of a pregnancy and others didn't.

Data based story-telling is another area where incidental data misuse may occur. Data based story-telling combines data with narrative, and can be used to make a compelling case to support specific goals. Dykes (2020, p. 39) reminds data science professionals that "in combination, data and narrative can form a formidable union, each strengthening the other in areas where it is weak. In a post-truth environment, we must be even more disciplined in how we craft and tell stories with data."

INTENTIONAL MISUSE

The intentional misuse of data is not limited to one organization or field (Davenport, 2020). Several professional sports teams have been accused of the intentional misuse of analytical data. Recent examples include the Houston Astros (stealing pitch signs from opponents) and the New England Patriots (illegal filming of opponents). The targeting of customers by the sub-prime mortgage industry and/or credit card companies is another example of unethical data use.

ETHICAL DATA USE & USERS

Since data misuse may be incidental, should the field of data science take a proactive ethics approach as suggested by Schmarzo? How exactly, would proactive ethics work in such a quickly expanding field? One approach is to provide guidelines for ethical data use. Stone (2020) suggests five criteria:

- Is the data authentic, trustworthy, and of known source?
- Will the data be used consistent with the original purpose of its collection?
- Did the data owners consent to this use of the data?
- How will the risk of unintended harm be addressed?
- Is the data free of bias?

Concerns with underlying bias in data have been expressed with facial recognition algorithms and recidivism predictions (Najibi, 2020; O'Neil, 2017). Data bias can arise from a number of sources (Lawton, 2020), including:

- Flawed training data due to underrepresented populations or collinear variables
- Presence of extreme outliers within data
- Incomplete data sets due to confirmation or selection bias

Even if data is unflawed, potential ethical issues persist because data ethics extends into every area and process of data science. General suggestions on preventing unethical workplace behavior have been provided by Brookins (2019):

- Establish a code of conduct, with specific expectations and consequences.
- Create a system of checks and balances.
- Owners and managers lead by example.
- Hire to match organizational values and show appreciation for employees.
- Bring in ethical experts for presentations / training.

Another proactive approach is the development and application of a code of ethics for data science. This approach is endorsed by Mitchell-Guthrie (2020), who suggests utilizing the code of ethics created for the Certified Analytics Professionals program (https://www.certifiedanalytics.org/ethics.php, presented in its entirety at the end of the paper). Mitchell-Guthrie argues that a code of ethics is needed because of the rapid growth and change in data science, its increasing omnipresence in daily life, the push for profit, and the potential harm that is possible from unrestrained data science.

In the same volume of essays, however, Cherry (2020) argues against the need for a code of ethics. Cherry argues for "regular and repeated doses of common sense" (p. 205) instead of codified ethical guidelines. Cherry bases this argument from the starting point that data, in and of itself, is neutral. Biases that may exist in the data are a result of human involvement and not of the data itself. Cherry argues that ethics lacks the context used in the application of common sense (note the term "context" was not present in the definition of ethics presented in the introduction of this paper) and as such, is not the best basis for guiding the actions of data science.

CONCLUSION

Improvements in technology and programming have led to an ever-increasing amount of data and analytical possibilities. Much good comes from data science, but as in any field where people are involved, there is also the potential for great harm. Whether data misuse is incidental or intentional, the outcome is still possibly harmful to innocents. Some members of data science push for a standard Code of Data Science Ethics while others argue just as strongly against the imposition of a rigid code.

In addition to defining data ethics, Floridi and Taddeo (2016) wrestled with the difficulty of balancing the possibilities of data science with the potential harm. The authors concluded that "Striking such a robust balance will not be an easy or simple task. But the alternative, failing to advance both the ethics and the science of data, would have regrettable consequences" (p. 2).

Has the ability to analyze data outpaced the growth of data ethics? As long as active discussions around analytics and ethics persist, the answer is probably no. If discussions do not result in specific conclusions or actions, however, the answer is most likely yes.

REFERENCES

Brookins, M. (2019, February 1). Ways to prevent unethical behavior in the workplace. *Houston Chronicle*. Retrieved from https://smallbusiness.chron.com/

Cherry, D. (2020). Data science does not need a code of ethics. In B. Franks (Ed.), *97 things about ethics everyone in data science should know* (pp. 204-205). Sebastopol, CA: O'Reilly Media Inc.

Davenport, T. (2020, March 4). The Houston Astros and the ethical use of data and analytics. *Forbes*. Retrieved from https://www.forbes.com/

Duhigg, C. (2012, February 16). How companies learn your secrets. *The New York Times*. Retrieved from https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Dykes, B. (2020). Data storytelling: the tipping point between fact and fiction. In B. Franks (Ed.), *97 things about ethics everyone in data science should know* (pp. 39-40). Sebastopol, CA: O'Reilly Media Inc.

Floridi, L., & Taddeo, M. (2016). What is data ethics? *Phil. Trans. R. Soc.* A 374: 20160360. https://royalsocietypublishing.org/doi/pdf/10.1098%2Frsta.2016.0360

Hand, D. J. (2018). Aspects of data ethics in a changing world: Where are we now? *Big Data, 6*(3), 176-190.

Lawton, G. (2020, October 26). 8 types of bias in data analysis and how to avoid them. Retrieved from https://searchbusinessanalytics.techtarget.com/

Mitchell-Guthrie, P. (2020). Ethical data science: Both art and science. In B. Franks (Ed.), *97 things about ethics everyone in data science should know* (pp. 218-219). Sebastopol, CA: O'Reilly Media Inc.

Najibi, A. (2020, October 24). *Racial discrimination in face recognition technology* [Blog]. Retrieved from https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

O'Neil, C. (2017). *Weapons of math destruction: How Big Data increases inequality and threatens democracy*. New York, NY: Broadway Books.

Price, D. (n.d.). *How much data is produced every day*? [Infographic]. Retrieved from https://cloudtweaks.com/2015/03/how-much-data-is-produced-every-day/

Siegel, E. (2020, October 23). When does predictive technology become unethical? *Harvard Business Review.* https://hbr.org/2020/10/when-does-predictive-technology-become-unethical

Schmarzo, B. (2020). Understanding passive versus proactive ethics. In B. Franks (Ed.), *97 things about ethics everyone in data science should know* (pp. 18-20). Sebastopol, CA: O'Reilly Media Inc.

Stone, S. (2020). Just because you could, should you? In B. Franks (Ed.), *97 things about ethics everyone in data science should know* (pp. 18-20). Sebastopol, CA: O'Reilly Media Inc.

Vuleta, B. (2021, January 28). *How much data is created every day?* [Blog]. Retrieved from https://seedscientific.com/how-much-data-is-created-every-day/

Webber, K., & Morn, J. (2018). *The uses and potential misuses of data*. Discussion session at AIR Forum 2018, Orlando, FL.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Kelly D. Smith
Central Piedmont Community College
kds.aewas@gmail.com
kelly.smith@cpcc.edu
www.linkedin.com/in/kelly-d-smith

CODE OF ETHICS/CONDUCT (Retrieved from https://www.certifiedanalytics.org/ethics.php)

INFORMS has developed the code of ethics/conduct for all Certified Analytics Professionals [see below]. All candidates and certificants participating in the certification process are required to agree to comply with the current and future provisions of this code.

Code of Ethics for Certified Analytics Professionals

Prepared by the INFORMS Certification Task Force

Background. The Institute for Operations Research and the Management Sciences (INFORMS) does not have an established code of ethics or guidelines for ethical practice that applies to the general membership. However, Article 1, Paragraph 2.v., of the INFORMS constitution states, "The Institute will strive to promote high professional standards and integrity in all work done in the field."

Applicability. This Code of Ethics applies specifically to those seeking (re)-certification as a Certified Analytics Professional (CAP®), but may be useful to other practitioners who use analytics. Clients, employers, researchers, policymakers, journalists, students and the public should expect analytical practice by CAP® certified individuals to be conducted in accordance with these guidelines. Application of these or any other ethical guidelines generally requires good judgment and common sense.

Purpose. This Code exists to clarify the ethical requirements that are important; inform the individual regarding rules and standards; hold the profession accountable; aid analytics professionals in making and communicating ethical decisions; help deter unethical behavior and promote self-regulation; and list possible violations, sanctions, and enforcement procedures.

General. Analytics professionals participate in analysis that aids decision makers in business, industry, academia, government, military, i.e. all facets of society; therefore, it is imperative to establish and project an ethical basis to perform their work responsibly. Furthermore, practitioners are encouraged to exercise "good professional citizenship" in order to improve the public climate for, understanding of, and respect for the use of analytics across its range of applications. In general, analytics professionals are obliged to conduct their professional activities responsibly, with particular attention to the values of consistency, respect for individuals, autonomy for all, integrity, justice, utility, and competence.

Responsibilities. This Code recognizes that analytics professionals have obligations to a variety of groups, including: society, employers and clients, colleagues, research subjects, INFORMS, and the profession in general. Responsibilities regarding each of these groups are further described below.

Society. All professionals have societal obligations to perform their work in a professional, competent, and ethical manner. Professionals should adhere to all applicable laws, regulations, and international covenants.

Employers and Clients. In general, it is the practitioner's responsibility to assure employers and clients that an analytical approach is suitable to their needs and resources, and include presenting the capabilities and limitations of analytical methods in addressing their problem. Analytics professionals should clearly state their qualifications and relevant experience. It is imperative to fulfill all commitments to employers and clients, guard any privileged information they provide unless required to disclose, and accept full responsibility for your performance. Where appropriate, present a client or employer with choices among valid alternative approaches that may vary in scope, cost, or precision. Apply analytical methods and procedures scientifically, without predetermining the outcome. Resist any pressure from employers and clients to produce a particular "result," regardless of its validity.

Colleagues. Analytics professionals have a responsibility to promote the effective and efficient use of analytical methods by all members of research teams and to respect the ethical obligations of members of other disciplines. When possible, professionals share nonproprietary data and methods with others; participate in peer review, focusing on the assessment of methods not individuals. Respect differing professional opinions while acknowledging the contributions and intellectual property of others. Those professionals involved in teaching or training students or junior analysts have a responsibility to instill in them an appreciation for the practical value of the concepts and methods they are learning. Those in leadership and decision-making roles should use professional qualifications with regard to analytic professionals' hiring, firing, promotion, work assignments, and other professional matters. Avoid harassment or discrimination based on professionally irrelevant bases such as race, color, ethnicity, gender, sexual orientation, national origin, age, religion, nationality, or disability.

Research Subjects. If a project involves research subjects, including census or survey respondents, an analytics professional will know and adhere to the appropriate rules for the protection of those human subjects. Be particularly aware of situations involving vulnerable populations that may be subject to special risks and may not be able to protect their own interests. This responsibility includes protecting the privacy and confidentiality of research subjects and data concerning them.

INFORMS and Profession. Analytics professionals will strive for relevance in all analyses. Each study or project should be based on a competent understanding of the subject-matter issues, appropriate analytical methods, and technical criteria to justify both the practical relevance of the study and the data to be used. Guard against the possibility that a predisposition by investigators or data providers might predetermine the analytic result. Remain current in constantly changing analytical methodology, as preferred methods from yesterday may be may be barely acceptable today and totally obsolete tomorrow. Disclose conflicts of interest, financial and otherwise, and resolve them. Provide only such expert testimony as you would be willing to have peer reviewed. Maintain personal responsibility for all work bearing your name; avoid undertaking work or coauthoring publications for which you would not want to acknowledge responsibility.

Alleged Misconduct. Certified Analytics Professionals will strive to avoid condoning or appearing to condone careless, incompetent, or unethical practices. Misconduct broadly includes all professional dishonesty, by commission or omission, and, within the realm of professional activities and expression, all harmful disrespect for people, unauthorized or illegal use of their intellectual and physical property, and unjustified detraction from the reputation of others. Recognize that differences of opinion and honest error do not constitute misconduct; they warrant discussion, but not accusation. Questionable scientific practices may or may not constitute misconduct, depending on their nature and the definition of misconduct used. Do not condone retaliation against or damage to the employability of those who responsibly call attention to possible scientific error or misconduct.

References.

Saul I. Gass, Ethical guidelines and codes in operations research, Omega 37(2009), 1044-1050.

American Statistical Association, Ethical Guidelines for Statistical Practice, August 7, 1999.

U.S. federal regulations regarding human subjects protection are contained in Title 45 of the Code of Federal Regulations, Chapter 46 (45 CFR 46).