Lies, Damned Lies, and Statistics: Deception and its Ethics as Applied to Statistics and Data Science

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Brendan Grimes

Summer 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Joshua Earle, Department of Engineering and Society

Intro

"You lied to me by telling the truth? . . . That's very good! May I use that?"

– Capt. Jack Sparrow, Pirates of the Caribbean: On Stranger Tides (2011)

During the 2020 Presidential Election Campaign, the Indianna GOP accused Mr. Pete Buttigieg, the then mayor of South Bend, Indianna of having one of the worst violent crime rates among American cities with populations greater than 100,000 (*Pete Buttigieg's South Bend,* 2019). This claim came from an USAToday article listing the worst cities in America based on the official FBI Uniform Crime Report's statistics (Stebbins & Comen, 2019). But as it happens, the GOP's claim was a smear feeding on USAToday's careless reporting. While the FBI's data did show an increase in violent crimes under Mayor Buttigieg's rule, and while the rate for violent crimes was indeed high enough to put South Bend on USAToday's list, as the New York Times pointed out, this was due to a reporting practice change surrounding aggravated assaults (Asher, 2019). The change caused assaults that the city had been classifying as simple to be classified instead as aggravated, making them count as violent crimes. Furthermore, the FBI noted this reporting practice change in a footnote to its statistics stating that the previous years' numbers were not comparable (FBI UCR, 2016).

This is an excellent example of perfectly accurate data being used carelessly and even maliciously to deceive others. Deceiving by using accurate but inappropriate data is only one of many ways that data is used to mislead. The New York Times article also takes issue with USAToday's treatment of St. Louis, Missouri in the same list. Although St. Louis does indeed have a very high rate of murders per 100,000 people, it also has a much smaller municipal area than most similar cities. In other similar cities, the more peaceful areas inside the city limits help to balance out the more violent areas. However, in St. Louis, the city boundaries are drawn too tightly around the more violent city center to allow this dilution to happen. This is an example of a failure to explain the context thoroughly leading to a misleading statistic. While this example likely is not malicious, it shows just how easy it is to utilize accurate data to arrive at a false or misleading conclusion.

Such data-based deception is all too commonplace in our society today. But researchers have shown that many people who deceive by strategically using true but misleading statements often feel as if they are not doing something so evil as lying outright (Rogers et al., 2017). On the other hand, those whom they deceive feel just as deceived as if the deceiver had in fact lied to them outright. Thus, there is haziness in people’s perceptions of the ethicality of deceptive or misleading actions.

In this paper, I seek to clarify the ethicality of deception in its various forms, particularly as applied to data science and statistics. To this end, I conduct an ethical analysis of deception using the moral teaching of the Catholic Church. Along with Rogers et al., Prümmer, and others, I focus on three broad categories of deception and discuss the ethicality of each. I then apply this understanding to the use of deception when working with data. Ultimately, I conclude that nearly all forms of deception, with very few exceptions, are morally reprehensible, and give examples of deceptive data use practices that fit the categories earlier identified.

Methods

Ethical Methods

I base the ethical aspect of my work on the moral tradition of the Catholic Church. I primarily use the *Handbook of Moral Theology* (Prümmer, 1921/2022), *The Catechism*

Explained (Spirago, 1899/2020), and *The Catechism of the Council of Trent*, also known as *The Roman Catechism*, issued by the order of Pope St. Pius V in the 16th century (1566/2021). For definitions, I draw upon the *Dictionary of Scholastic Philosophy* (Wuellner, 1956/2012).[1]

The Catholic Moral Tradition is a very large subject. But although it is quite thorough and complex, it is entirely contained in the Ten Commandments, albeit in a greatly compressed format. Our Lord himself, when asked which commandment is the greatest, summarizes the entirety of Catholic morality and teaching saying "Thou shalt love the Lord thy God with thy whole heart, and with thy whole soul, and with thy whole mind . . . [and t]hou shalt love thy neighbor as thyself. On these two commandments dependeth the whole law" (*Douay-Rheims Bible,* 1941/2018, Matt. 22:37-40). Throughout the centuries since the coming of Christ, the Church has continuously sought to explain and clarify all the implications and subtleties that can be drawn from the Ten Commandments. Hence, there is much literature on the subject, despite the relative simplicity of its origins. In the case of the discussion of deception, I am concerned primarily with the 8th Commandment: "Thou shalt not bear false witness against thy neighbor" (*Douay-Rheims Bible,* 1941/2018, Ex. 20:16) which is understood by the Church to prohibit all forms of lying and deception (*The Roman Catechism*, 1566/2021).

To a certain extent, the Catholic Moral Tradition can be understood as a form of Virtue Ethics. Indeed, Aristotle's works have been utilized quite effectively within the Catholic Moral Tradition, beginning with the work of St. Thomas Aquinas. However, unlike the variety of Virtue Ethics where different societies and cultures may hold different things to be virtuous, the Catholic Moral Tradition has sixty-four commonly identified and well-defined virtues that are

[1] Scholastic Philosophy is "the systematic philosophy developed in the Middle Ages from Aristotelian and Augustinian roots [meaning St. Augustine of Hippo], highly developed by St. Thomas Aquinas" (Wuellner, 1956/2012, p. 112). It is generally considered to be the main philosophical system used by the Catholic Church.

universal in time and space.[2] Catholic Moral Tradition also differs from the common understanding of Virtue Ethics in that morality is not determined by an individual's thought experiment of considering what a virtuous person would do, although this is still considered a highly useful exercise. Rather, morality is determined by whether one is obedient to God and His commandments. In light of this, the virtues are a set of good habits, the perfect practice of all of which guarantee one is perfectly obedient to God. Acting contrary to the virtues is typically immoral but is always at least defective. In and of itself, deception violates the virtue of truthfulness, but depending on the circumstances, deception may violate other virtues as well (Prümmer, 1921/2022). For example, falsely speaking ill of another would also be contrary to the virtues of justice and charity.

Additional Resources

While ethical analysis is the center point of my work, in applying this analysis to the topic of data, I have used the works of several scholars on various related topics, especially to furnish examples related to deception in data-science and statistics. Lauren Willis, in her 2020 article "Deception by Design" gives an excellent overview of deception in the realm of digital marketing, which often overlaps with data-related activities. The study by Lewis et al. (2017) on AI bargaining systems provides a fascinating example of a machine learning to deceive of its own accord. Cathy O'Neal's *Weapons of Math Destruction* (2016) also provided excellent examples of deception in data science and statistics. The work of Corple and Linabary in "From data points to people: feminist situated ethics in online big data research" (2020) was useful in highlighting the problems that arise from ignoring the context of data.

[2] A list of these can be found at https://sensustraditionis.org/Virtues.pdf

One of the most helpful resources for my discussion was “Artful Paltering: The Risks and Rewards of Using Truthful Statements to Mislead Others” by Rogers et al. (2017). In this article, they discuss a somewhat more obscure form of deception, which they term “paltering”. As made clear by the title, they define paltering to be “Using Truthful Statements to Mislead Others”. This concept of paltering contains both what Prümmer (1921/2022) calls “Mental Reservation” (to mean something in a sense other than that which is commonly understood) and what Prümmer calls “Amphibology” (a statement that could have multiple meanings) (sec. 293). It is particularly relevant when discussing data and its abuses since, apart from falsification of data or omission of relevant data points, most deception involving data is of this more subtle flavor. Furthermore, while understanding the more blatant deceptions connected with data and why they are wrong is useful, understanding why it is wrong to palter is more useful because it is less obvious. While Rogers et al. do not discuss the ethical reason why paltering is wrong, by comparing their treatment with the treatment of related topics in Prümmer and Spirago, we can understand its immorality with relative ease.

Ethical Analysis of Deception

Rogers et al. (2017) identify three separate flavors of deception: lying by commission (“the active use of false statements”) lying by omission (“the passive omission of relevant information”) and paltering (“active use of truthful statements to convey a misleading impression”) (p. 456). In the following analysis, I first discuss why deception is immoral, and then discuss what exceptions to this exist. In the subsequent section, I examine how these three categories of deception can be applied to topics surrounding data science and statistics.

Why Deception is Immoral

The Catechism of the Catholic Church (1566/2021) states that the 8th Commandment forbids lying when it condemns false witness, mentioning that lying is also plainly condemned in many other passages of Scripture besides the Ten Commandments. Thus, since God has condemned it, it is clearly immoral. However, as with all God's laws, there is a good reason why He forbids deception. Prümmer gives four primary reasons for the inherent immorality of lying:

> The intrinsic reason for the evil character of lying is that it is opposed to: *a)* the natural purpose of speech which is given to man to reveal what is in his mind; *b)* natural human intercourse[3] which is disturbed by lying; *c)* the good of the listener who is deceived by the lie; *d)* the welfare of the speaker himself who, although he may obtain some temporary advantage from the lie, will suffer greater evils in consequence (Prümmer, 1921/2022, sec. 292).

While the first reason is particular to the vocal nature of lying, it could be generalized by replacing "speech" with "communication" to incorporate other forms of deception. The second two reasons are readily applicable to written or digital forms of deception. Though it is easy enough to understand the fourth reason by considering the punishments for evil actions rendered by God after death, the fourth reason is also understandable in a purely earthly sense. Violence is defined as "action contrary to the nature of a thing" (Wuellner, 1956/2012, p. 131). Thus, since deception is contrary to the nature of our powers of communication, properly speaking, deception is doing violence to oneself, even if the deception is meant to bring about something good.

[3] In the primary sense of the term, not as in sexual intercourse

Prümmer (1921/2022) also states that “no reason whatsoever can justify [a lie]” (sec. 292). Though this may seem somewhat strict, it stems from the inherently evil nature of lying. Notably, however, the act of mental reservation is in certain cases permissible, as is amphibology. These cases are treated in the next section.

Forms of Deception

To lie by commission or by omission, as defined by Rogers et al. (2017) is lying strictly so called. As such, for the reasons already stated, it is always and unconditionally wrong. Spirago (1899/2020) elaborates upon this even further saying “Not even to save one’s own life or the life of another, is a falsehood justifiable” (p. 415). Spirago (1899/2020) does point out that clearly false jests and ambiguous (but not false) answers to questions which the asker has no right to know do not fall under this requirement (p. 415). This is because they are not lies properly speaking, as in the first case the action is not deceptive, and in the second case, no falsehood was uttered, and though the asker may be deceived, it is not a matter which he should be permitted to know.

The second exception mentioned by Spirago is an example of what Prümmer (1921/2022) calls amphibology (sec. 293). Prümmer (1921/2022) also explains that it is permissible with sufficient reason to use amphibology and/or broad mental reservation, that is, to “restrict the sense of the words used to a meaning different from their obvious meaning” in such a way that “a prudent man could gather the intended meaning from the surrounding circumstances” (sec. 293). If it is not done in this way, it is termed strict mental reservation and is entirely prohibited as it is then a lie. Prümmer (1921/2022) shows that broad mental reservation and amphibology are concealment of truth without uttering falsehood, and as such, when the occasion requires it, are permissible. These concepts of mental reservation and amphibology are both contained in the concept of paltering defined by Rogers et al.

Ethics of Deception applied to Data Science

In the previous section I discussed the various forms of deception and the reasons for their immorality. Next, I discuss how this can be applied to the topics of Data Science and statistics. Data Science is here used to mean those aspects of science that rely on a non-negligible quantity of data. I recognize that this is common to most empirical science, however, I use the term Data Science because I am concerned *only* with the aspects of empirical sciences that have to do with the use of data, and not those that have to do with any specifics of one field or another.

Falsifying data is lying by commission. It is deliberate and intentional, and its purpose is to produce a result contrary to that which would have otherwise been obtained. In other words, it is communicating a deliberately false conclusion. However, using false data unintentionally is not the same thing morally. While this would still be a false conclusion, if it were done truly without the scientist's knowledge, since it lacks intention, it could not be properly called a lie of commission. Rather, it would either be negligence or without blame entirely.

One of the most deeply deceptive uses of data science are what Cathy O'Neal (2016) calls Weapons of Math Destruction, or WMDs for short. These are mathematical models that are based on data sets that are statistically too small to be correct, lack a feedback system for correcting the model's errors, cause non-negligible harm, and have substantial scale. The foundational issue though, is that they are based on input data that is statistically unsupportable for providing the conclusions that the models draw and that they lack the feedback mechanisms by which this could possibly be corrected. However, they are often treated as being objective, or accurate, or at least better than a human being in terms of objectiveness and correctness. Thus, these WMDs are a form of lying by commission because they are marketed as an objective and accurate model despite being statistically invalid.

While lying by commission is perhaps the easiest form of deception to perceive, lies of omission are also prevalent in data science and statistics. Selectively choosing only data which supports one’s hypothesis and omitting that which contradicts it such that the outcome is changed is a lie by omission. It is also a lie of omission to ignore the context of the data to such a degree that the result is affected. Intentionally failing to communicate a sufficient degree of context for the target audience such that they are left with a false conclusion is an example of this. Corple and Linaberry (2020) warn that disembodiment of data, i.e., ignoring the context from which it is drawn, leads to disproportionately adverse consequences for those who are already marginalized. However, unintentional failure to provide context is not immoral, so long as due diligence on the author’s part is observed.

Paltering is quite commonplace in the manipulation of visual data presentation elements such as graphs. For example, distorting graph axes on a bar chart of vote count percentages to make a very close election race look as if one candidate or another is far in the lead. The author in this example is strategically presenting the data to give the viewers the impression that one candidate is winning by a substantial margin, even though in reality, the data show that the race is quite close. There is no false information, nor are there any missing data, so long as the actual numbers are posted on the graph. Hence, it is not lying in either form. However, despite only using true and accurate information, the presenter is insinuating a false conclusion. Though this is clearly an instance of paltering, its ethicality merits further discussion.

As previously noted, paltering can be broken down into mental reservation and amphibology. The previous example, though it is deceptive, could be classified as mental reservation because it uses the normal way that people interpret the visual cues of a graph to evoke a conclusion other than that which the chart is really showing. It is more properly broad mental reservation, because, though a false impression is being evoked, a prudent man could read the numbers next to the graph bars (assuming these are included) and form the correct conclusion with relative ease. Although broad mental reservation is not condemned in all cases, Prümmer (1921/2022) specifies that it should be used “when there is need or sufficient cause” rather than whenever one pleases (sec. 293). Thus, if it were done without a good reason, the previous example would still be considered immoral despite being broad mental reservation.

Strict mental reservation is also possible in data-related fields. Continuing with the previous example, if the graph had been presented without sufficient numerical information to distinguish that the race was close, then it would be strict mental reservation. Although the graph would still be technically accurate, the presenter would be showing it in such a way that a prudent man is unable to distinguish the actual meaning and is left only with the false conclusion. As Prümmer (1921/2022) points out, strict mental reservation is always immoral (sec. 293).

Willis (2020) gives many examples of mental reservation in digital marketing. One such example is strategic placement of content on webpages such that common consumer reading patterns will pass over the downsides of an item, despite the presence of this information on the page (Willis, 2020). Consumers are thus manipulated into forming false impressions about the item via their reliance on a common understanding of the structure of webpages. Though this is not directly related to data or statistics, often, these marketing tricks are performed real time by machine learning models trained on the behavior and patterns of users. Willis (2020) shows that, if left unchecked to maximize sales, data-driven marketing models will use deceptive practices, whether their creators intend this or not. Willis (2020) points to the study conducted by Lewis et al. as an excellent example of this reality. In that study, they discuss their development of a bargaining AI that learned, entirely on its own, to deceive (Lewis et al., 2017). Thus, though they may have no intention to deceive, the creators and users of such systems must actively prevent them from resorting to deceptive practices.

Amphibology relies on ambiguity of meaning in the information given. This contrasts with mental reservation, which preys upon the way that people commonly understand things. However, though they can be used to manipulate people's opinions, and often are used to create false impressions, statistical measures and data-based information typically are not truly ambiguous. This is because ambiguous information does not inform (neither truly nor falsely) and thus is not really information. If one cannot interpret the meaning of a graph at all, it typically is considered useless, rather than deceptive, likewise for an inconclusive statistical model. However, despite their scarcity in statistical and data science contexts, as with broad mental reservation, instances of amphibology are moral if they are done with sufficient cause, and neither used lightly nor for malicious purposes.

Conclusion

Deception in data science and statistics is unfortunately all too commonplace. While there are many good scientists and statisticians that strive to always avoid misleading those who read or use their works, there are also many who choose, for whatever reason, to deceive their fellow men. Particularly, the WMD pattern identified by O'Neal (2016) is strikingly common, with many such models in existence across many disciplines, from e-commerce to personnel management. But before we can address these sorts of issues in society, we must understand what forms the problem can take and what makes them problematic. Thus, a precise and specific knowledge of the ethics behind deception is essential to solving the crises of data deception that harass the modern world.

Though irresponsible or evil use of these technologies can cause great harm, their proper use can be a source of much good. Modern data technologies have enabled amazing advances in many fields, from fraud detection to medical imaging. Technologies influence society only because of how we choose to use or not use them, though their design is often meant to shape how they are used. Thus, we must never forget the importance of human choices, and therefore ethics, when studying how technology harms or helps society. There will always be liars, and statistics will always be one of their tools, but perhaps we can hope for a future in which the systematic data-based deceptions we see today are no longer quite so commonplace.

References

Asher, J. (2019, December 17). South Bend and St. Louis, Where Crime Statistics Can Mislead. *The New York Times*. https://www.nytimes.com/2019/12/17/upshot/crime-statistics-south-bend-st-louis-misleading.html

Corple, D. J., & Linabary, J. R. (2020). From data points to people: Feminist situated ethics in online big data research. *International Journal of Social Research Methodology*, *23*(2), 155–168. https://doi.org/10.1080/13645579.2019.1649832

FBI. (2016). *FBI Uniform Crime Report: 2016, table 6: Violent crime, Indianna cities*. FBI. https://ucr.fbi.gov/crime-in-the-u.s/2016/crime-in-the-u.s.-2016/tables/table-6/table-6-state-cuts/indiana.xls

Lewis, M., Yarats, D., Dauphin, Y. N., Parikh, D., & Batra, D. (2017). *Deal or No Deal? End-to-End Learning for Negotiation Dialogues* (arXiv:1706.05125). arXiv. http://arxiv.org/abs/1706.05125

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Internet materials). Crown. https://proxy1.library.virginia.edu/login?url=https://ebookcentral.proquest.com/lib/UVA/detail.action?docID=6108230

Pete Buttigieg's South Bend: "One of the Most Dangerous Cities" in America. (2019, April 11). The Indiana Republican Party. https://www.indiana.gop/news/pete-buttigieg%E2%80%99s-south-bend-%E2%80%9Cone-most-dangerous-cities%E2%80%9D-america

Prümmer, D. (2022). *Handbook of Moral Theology* (G. Shelton, Trans.; 1st ed.). Benedictus Books. (Original work published 1921, first English translation 1956)

Rogers, T., Zeckhauser, R., Gino, F., Norton, M. I., & Schweitzer, M. E. (2017). Artful paltering: The risks and rewards of using truthful statements to mislead others. *Journal of Personality and Social Psychology*, *112*(3), 456–473. https://doi.org/10.1037/pspi0000081.supp

Spirago, F. (2020). *The Catechism Explained An Exhaustive Exposition of the Christian Religion* (R. Clarke, Ed.). Mediatrix Press. (Original work published no later than 1899)

Stebbins, S., & Comen, E. (2019, February 1). *These are the worst cities to live in based on quality of life*. USA TODAY. https://www.usatoday.com/story/money/economy/2018/06/13/50-worst-cities-to-live-in/35909271/

The Catechism of Council of Trent (J. McHugh & C. Callan, Trans.). (2021). Preserving Christian Publications, Inc. (Originally Published in Latin 1566, this translation in 1923)

The Holy Spirit. (2018). *The Holy Bible* (English College at Douay (Old Testament, 1609) & English College at Rheims (New Testament, 1582), Trans.). Loreto Publications. (This version originally published 1941).

Willis, L. E. (2020). Deception by Design. *Harvard Journal of Law & Technology*, *34*(1), 115–190.

Wuellner, B. (2012). *Dictionary of Scholastic Philosophy*. Loreto Publications (Original work published 1956).