A 3D-rendered illustration of modern machine learning with neural network and coding. (Image by Adobe Stock)

The Center of Gravity in Artificial Intelligence Ethics Is the Dataset

Capt. Timothy Naudet, U.S. Army

Capt. Robert B. Skinker, U.S. Army

Equitable outcomes must exist for every artificial intelligence (AI) use case, from topics as sensitive as racial demographics to topics as tactically paramount as visually detecting enemy vehicles. An AI model achieves ethical outcomes through the datasets used in its construction. This article relates AI ethics to an intuitive application of AI: using a drone to visually identify and engage targets. Building an equitable database increases the probability of the desired outcome: successful drone engagements. The "actionable insight" from this article is to ensure that AI models use equitable dataset engineering; every military AI dataset must be equitably balanced.

The dataset is the center of gravity in AI ethics because a human ultimately decides what information goes into the dataset and how it is organized. A human 'trains' the model on the dataset and determines whether the resultant model equitably estimates outcomes on previously unseen test data. The model's estimations, often called predictions, reflect the biases inherent in the assembled data.

AI is a broad term that encompasses machine-learning models. The estimated outputs from machine-learning models are strongly dependent on the datasets involved, and the ethics surrounding machine learning and its applications are subsequently dependent on those same datasets.. AI ethics should be evaluated on the use of explicit steps in engineering the dataset such as ensuring unbiased sampling, proper acquisition, consent, license, approval, and an equitable outcome. Engineering ethical guidelines will reduce ethical failures when applying machine learning.

Dataset Balance and Fairness

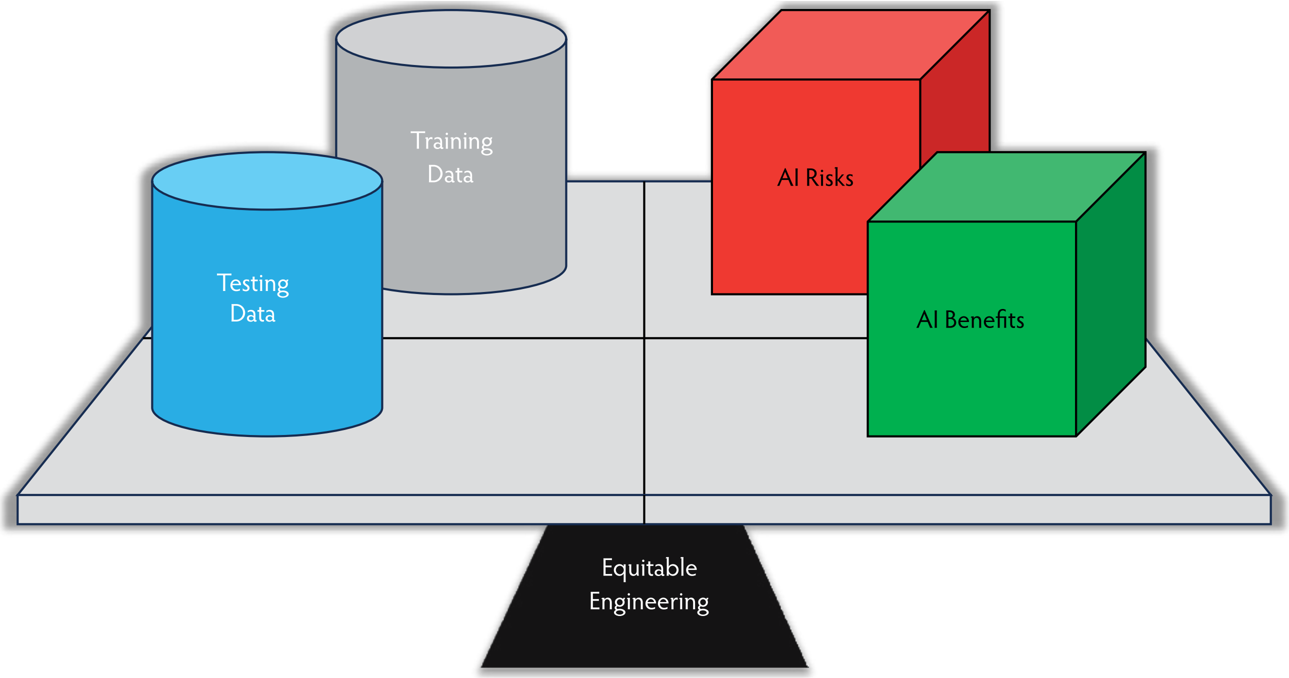
The center of gravity in AI ethics resides in the dataset. This article does not address the social ethics involved in the application of AI. The decision to use AI in self-driving cars, warfare, or recidivism are out of scope. This article is exclusively concerned with datasets.

The center of gravity is positioned as such because a human ultimately *decides* if the dataset is an acceptable set of information from which to build a machine-learning model. Since AI models are trained on the information contained in the human-selected datasets, the decision to use these datasets propagates the human bias incorporated during their assembly. Bias propagation continues to the model's estimations, consequentially affecting the benefits and risks of the model. Achieving an ethical use of artificial intelligence requires deliberate, professional effort to balance a dataset for an equitable outcome.[1]

Immediate, relevant examples of military AI posing ethical concerns are those in the Ukraine and Gaza conflicts. Israel is using AI to produce target reports for indirect fires and both sides of the Ukraine conflict are using autonomous loitering munitions.[2] There do exist complications in implementing these technologies, such as the widespread use of counter-unmanned aircraft system electronic warfare, but these are outside the scope of this article.[3] The focus is how these AI systems are functioning as compared to their *trained* use. A 2021 report produced by the Jewish Institute for National Security of America on the 2021 Gazan conflict discussed the advantages of the Israeli targeting AI, referred to as "Gospel" in English. This AI is the same one used in the ongoing Gaza conflict that began in October, 2023. The most notable advantage was the unmatched ability to process data and recommend targets. Gospel proved to be fifty times faster than the conventional human-analyst targeting system. However, a critical ethical complication arose from the lack of equitable dataset engineering.[4] The Jewish Institute for National Security of America reported

Capt. Timothy J. Naudet, U.S. Army, serves as a data scientist for the Artificial Intelligence Integration Center, Army Futures Command in Pittsburgh. He commissioned in 2017 from the U.S. Military Academy with a BS in chemistry and later graduated Carnegie Mellon University with a Master of Information System Management–Business Intelligence and Data Analytics in 2023 as a scholar of the Artificial Intelligence Scholars Program.

Capt. Robert B. Skinker, U.S. Army, is an assistant operations officer for 5th Engineer Battalion, 36th Engineer Brigade. He was commissioned in 2016 from Virginia Tech with a BS in mechanical engineering and later graduated Carnegie Mellon University with a Master of Information System Management–Business Intelligence and Data Analytics in 2023 as a scholar of the Artificial Intelligence Scholars Program.



(Figure by Capt. Timothy J. Naudet)

Balancing Act

This demonstrates the four-way balance that occurs among testing data, training data, AI benefits, and AI risks. An imbalance in any of the four components will produce an unethical AI model. AI risks and AI benefits refer to using the AI model's predictions.

that although "[Gospel] had plenty of training data for what constituted a target, it lacked data on things that human analysts had decided were not targets."[5] Despite being a well-developed AI targeting model, data which contained no targets was not used in the training dataset. Consequently, Gospel was unable to learn what *not* to identify. The "non-target" identifying skill might have been easily overlooked as the skill is commonplace in human targeting analysts, but the source of the issue, the bias in human sourced datasets, rendered Gospel an incomplete model. This bias might be related to the disparity of target types required for Gospel to detect, but a lack of further details prevents this analysis. AI models must be trained and tested on data that both does and does not contain the pertinent task. The information that is absent is as equally important as the information that is present.

The importance of absent data is reinforced by AI used in self-driving cars. A popular issue in self-driving cars includes *edge cases*. To put it simply, AI doesn't understand certain situations, such as a person climbing out of a manhole in the street.[6] The prevalence of different edge cases in driving renders a driving model vulnerable to failure. Lex Fridman et al. provide an incredibly thorough analysis of automated driving in comparison to human driving.[7] Their research addresses the parameters involved in how to *model* driving. Sorin Grigorescu et al. provide an extensive mathematical review of approaches in automated driving and present a lengthy discussion on the datasets involved.[8] Grigorescu et al. describe in great detail the marriage of sensors, data, and algorithms that influence a model's predictions. Both studies address edge cases incorporated into the datasets used in AI model development. Both address training a model to *generalize* predictions for optimal performance for edge cases. This generalization is a result of meticulously engineered real world driving datasets. Datasets drive AI. Equitable dataset engineering is critical for an ethical AI.

The Ukrainian conflict experienced AI use in both information processing as well as unmanned aircraft system (UAS) target identification.[9] Of interest is a Ukrainian AI that can detect military vehicles in camouflage. The ability to detect vehicles in camouflage initially seems commonsense. Militaries use camouflage, so this should be a basic requirement. However, the implicit data collection methods illuminate importance. Collecting camouflaged military vehicle data—where Ukraine likely doesn't have the resources to stage equipment for rehearsal data collections—indicates their data might be actual combat footage. The point is that the data used to train the model is highly relevant. This relevance has the potential to include data on what is not a target. There is likely high fidelity in the dataset. Compared to Gospel, the parameters of the Ukrainian AI might be more limited—it might be only required to detect military vehicles, which reduces the range of potential errors. There of course might exist other targets, such as infrastructure, but deeper analysis of the Ukrainian target detection is outside of this article.[10]

Outside of military applications, ethical artificial intelligence concerns often reside with the societal impacts of the model's predictions. These concerns rightfully include the mistreatment of minority demographics, where minority encompasses race, sex, religion, etc. Further concerns include unemployment that accompanies a new technology as well as potential violations of rights such as encroaching on privacy when collecting data.[11] These concerns are outside the scope of this work, but regardless of AI use case or negative impact, every artificial intelligence shares one common architecture component: a dataset.

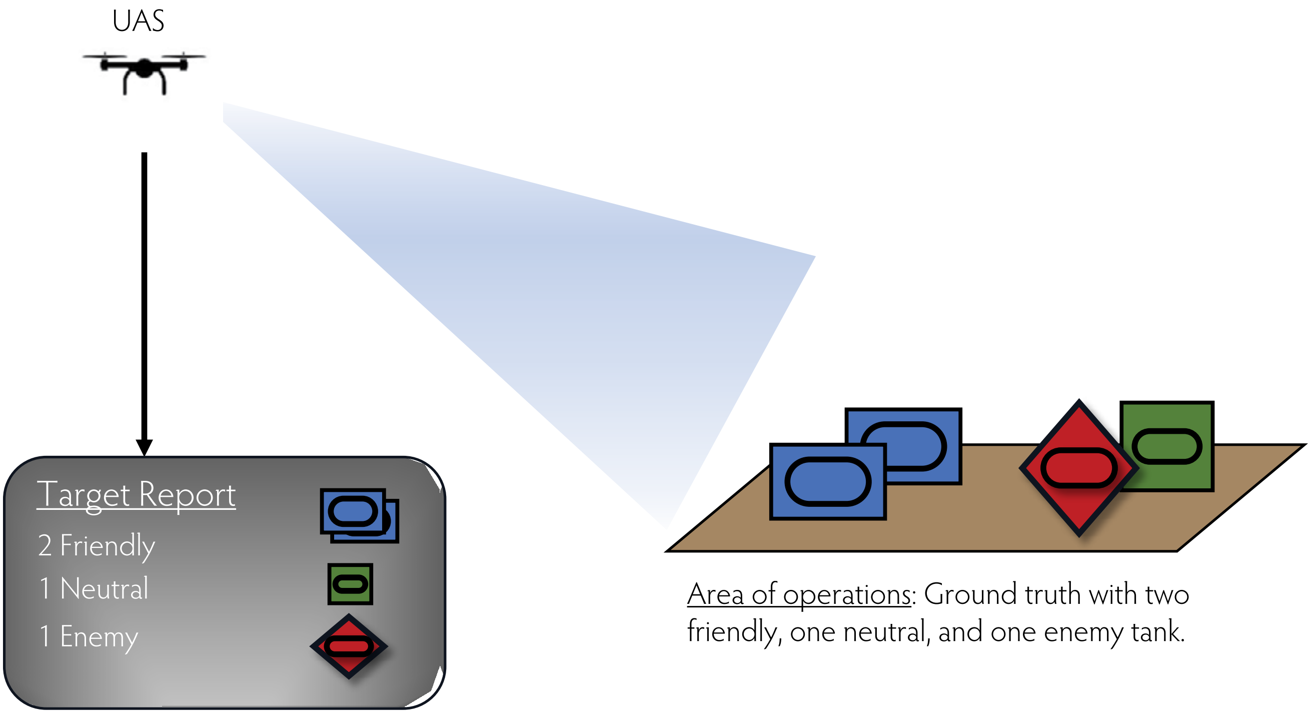
Stuart Russell and Peter Norvig provide a list of common AI ethical principles that are used to ensure technology contributes to "good" outcomes.[12] There also exists communities for ethical operations in machine learning and computing. The Institute of Electrical and Electronics Engineers (IEEE) has core values for the implementation of technology.[13] The IEEE completed fantastic work with "P7001: Proposed Standard of Transparency."[14] Alan F. T. Winfield et al. discuss different stakeholders, levels of transparency, and overall ethics of autonomous systems. They mention that for "learning systems, [transparency] includes details of the composition and provenance of training data sets."[15] Learning systems are entirely dependent on their data, and their dataset should be equitably engineered. Equitable dataset engineering should be an explicit measure of ethical AI.

Datasets can be as variable in the digital data type (video images, large time series, text, and tabular database) as their place of origin (battlefields, commercial roadways, universities, stock markets, laboratories, or city and rural environments). These differences may be subtle or obvious, but all the factors must be weighed toward equitable ethics. Models are not independent from their datasets. For example, an exclusive computer vision model would not use a dataset from an exclusive natural language processing model. The models and datasets are not interchangeable. However, every type of AI begins with a dataset. This ubiquity commands the imperative for AI ethics to start at the dataset.

Understanding the importance of the dataset is best accomplished by examining AI model predictions with an intuitive task called *image recognition*. The target recognition used in Ukraine enables an easy way to understand complications in datasets or models (refer to figure 1).

Figure 1 demonstrates expected target recognition when training data is complete. However, if an AI model is intended to detect objects in an image, but some desired objects are not in the dataset used to train the model, the AI would be unable to detect those objects.[16] The performance of this model would be poor. This performance is roughly explained by the overarching programming principle "garbage in, garbage out."[17] The output of a program depends on the input. Poor input equals poor output; biased input leads to biased output.

Facial recognition, a use case in computer vision, has obvious ethical complications if a model does not function properly. Valeriia Cherepanova et al. provide a clear escalation of facial recognition ethics when they state, "Incorrectly tagging a personal photo may be a mild inconvenience, but incorrectly tagging the subject of a surveillance image could have life changing consequences."[18] Proper function in facial recognition requires designing the system to maximize accuracy and reduce error. A component of proper function includes fairly detecting every demographic of person, with special care afforded to the minority demographics previously mentioned. Generally speaking, and without specific technical metrics, the crucial understanding of ethics in facial recognition is to ensure each unique demographic has an equivalent *accuracy* of detection. The goal is to make the model predictions equitable.[19] The goal is *not* to detect minority demographics in equal proportion to the majority demographic. The minority demographics are often underrepresented in


UAS
Target Report
2 Friendly
1 Neutral
1 Enemy
Area of operations: Ground truth with two friendly, one neutral, and one enemy tank.

(Figure by Capt. Timothy J. Naudet)

Figure 1. UAS Target Detection

The target report from the UAS matches the ground truth in the area of operations. This represents a proper functioning model.

the datasets, but they deserve an equitable detection probability when compared to the detection probability of the majority demographic.

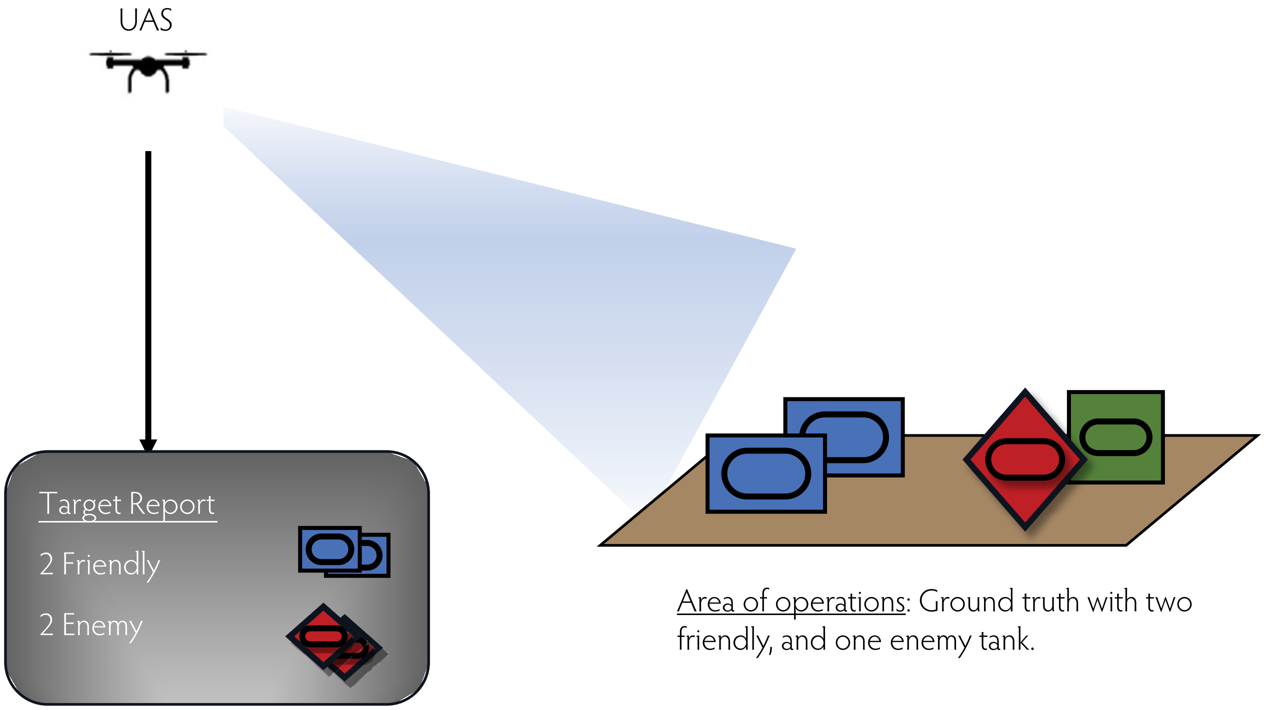
Despite the disparity in ethics between target detection and facial recognition, both ethical issues originate from the datasets. Fairness and bias in machine learning stem directly from the dataset.[20] The model will learn from the training data and will apply learned assumptions to new data.[21] If the dataset does not have an appropriate balance of vehicles or human faces to ensure fair detection, the model will not have acceptable accuracy or ethics. The performance and ethics of the model both independently rely on the "garbage in, garbage out" principle. Datasets drive AI, and consequently, the decision to use an imbalanced dataset is unethical. AI models should not reflect societal biases, and these biases are diminished through ethical dataset engineering.

The potential military implications of an incomplete dataset are obvious. Errors in detection could lead to unethically engaged tanks. Consider figure 2, where the model misclassifies the neutral tank as an enemy tank.

This figure demonstrates ethical concerns if an incomplete data set were to be used. The missing data in this case refers to a neutral armor formation. A neutral tank *might* be misclassified as an enemy tank.

A real-world example of potential ethical hazard is Israel's Harop missile.[22] As a loitering asset, the Harop missile is designed to identify a target based on an engagement criterion and engage without a human in the loop. Note that a human-in-the-loop option exists and is explicitly encouraged as an operational option. This option allows a human to abort missions to reduce collateral damage, but it is not required for the Harop to engage. The current criterion is communication signals, but this could easily be converted to visual detection.

Ethical concerns about a model's prediction must begin with the dataset used to train the model. Should data be missing from the dataset, the *owners* must introduce more data to the dataset to prevent unethical model outputs. The converse is also unethical. Should a dataset contain too many samples from a single demographic, the model will propagate biases inherent


UAS
Target Report
2 Friendly
2 Enemy
Area of operations: Ground truth with two friendly, and one enemy tank.

(Figure by Capt. Timothy J. Naudet)

Figure 2. Improperly Trained UAS Target Recognition

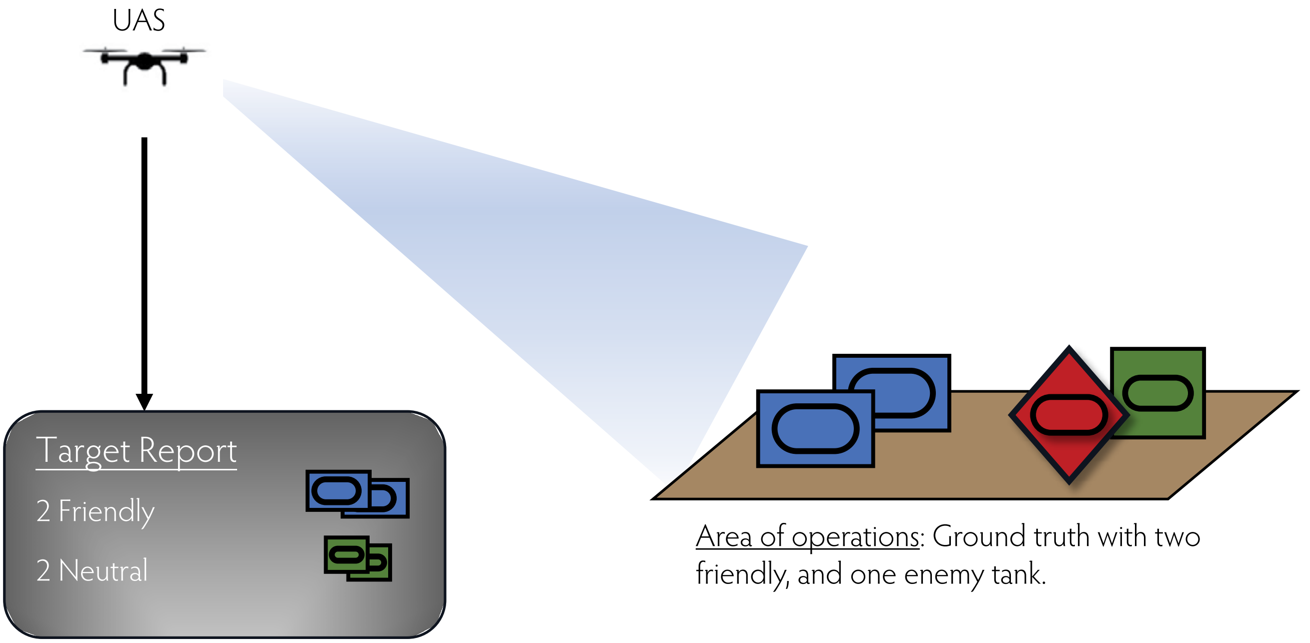
This figure demonstrates a failure in UAS target recognition. When provided an area of operations, a UAS correctly detects two friendly tanks but erroneously detects two enemy tanks.

in the dominant samples. Returning to the military tank recognition example, consider figure 3, where an overcorrection of neutral tank data may again lead to improperly labeled targets. This overloaded imbalance might produce biases equivalent to those produced through absent data.

Ethics in artificial intelligence must begin by curating robust and complete datasets. This effort must establish as much training data and sample disparity as possible.[23] *Robust* refers to training data that encompasses a wide range of data in both sample type and sample quality. *Complete* refers to training data that encompasses complex data. Robust data helps improve the model's true predictions while complete data helps reduce the model's false predictions.

The testing dataset is just as critical as the training dataset. Without delving into technical details, the test set is the data withheld from training and used to affirm the performance of the model after training. Data from the training and testing sets are exclusive. The test data must remain hidden from training the model. These datasets must never mix. The test set must be equally well engineered to ensure the model does not propagate bias onto unseen data. Cherepanova et al. explore the importance of testing set bias in facial recognition. Their results show an antagonistic relationship might develop between the two datasets, where the test set retains inequivalent proportions of demographics subsequently resulting in a misrepresentation of performance.[24]

After knowing the requirements for an ethical dataset, an obvious issue becomes specific methods of balancing the dataset—determining what constitutes "equitable." These determinations are difficult because unlike armored vehicles in an area of operations, there is often no *ground truth* for ethical comparison. Perhaps the best example that demonstrates the complexities in dataset balance is predicting recidivism using the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS).[25]


UAS
Target Report
2 Friendly
2 Neutral
Area of operations: Ground truth with two friendly, and one enemy tank.

(Figure by Capt. Timothy J. Naudet)

Figure 3. Improper Target Recognition

This figure demonstrates an abundance of neutral tank data and the consequential misclassification of the enemy.

The COMPAS model is used by criminal courts to predict the probability of recidivism and violent recidivism.[26] Understanding complexities in the model's predictions requires understanding bias inherent in the dataset. Recidivism data has inherent flaws because every criminal dataset will reflect those who were *convicted*, not those who were actual *criminals*.[27] This misrepresentation, even if not intentional, will propagate the bias in the conviction data to the model's predictions. Russell and Norvig establish the imperative for ethical data management and reducing impact of bias by stating, "First, understand the limits of the data you are using."[28]

Datasets must be equitably engineered within the intended use case. Although it is impossible to maximize all aspects of fairness, the first step in assembling a dataset is "to decide what counts as fair."[29] An inexhaustive ethical list of objectives include the following: equal opportunity, demographic parity, and equal impact.[30] The context of each unique artificial intelligence project will determine the extent of pursuit for these objectives, but assembling an ethical dataset for ethical prediction is the common goal.

Balancing a dataset to achieve equitable predictions is often challenging and counterintuitive. Cherepanova et al. provide wonderful insight into these complications. In the context of adding data to balance prediction accuracy for a single demographic group, they reveal that "overrepresenting the target demographic group can sometimes hurt the group."[31] The authors illuminate that bias in facial recognition can be concealed when using random selections for training and testing data even when following standard, randomized splitting procedures. They demonstrate that the subjects of prediction—such as male versus female—and the dataset's entire composition may equally contribute to model bias. Ultimately, achieving a balanced dataset for an equitable outcome requires close examination of the model's predictions and focused effort to reduce biased predictions.

Curating Datasets

Having discussed the requirements for a professional, ethical dataset, the logical next step is to discuss the assembly of data. Engineering a professional dataset is laborious. Significant attention is required to collect, label, and organize a dataset. The labor involved must be orchestrated with the final product in mind: the ethical predictions. This will prevent the owners from accumulating technical debt that is more challenging to correct after the fact. Capturing robust and complete data in the engineering phase will encourage better dataset balance following pruning the dataset and adding sample disparity.

Fortunately, the internet is the greatest contributor to available data sources. It receives up to 10^{18} bytes of new data per day.[32] YouTube alone provides up to three hundred hours of new data every minute.[33] Many standardized datasets are available for download across all disciplines of machine learning.[34] The advent of data is so prominent that "data drives the operation; it is not the programmers anymore, but the data itself that defines what to do next."[35] The overwhelming amount of information provides the foundation for interesting machine learning research and expansive fields of study.

Internet data prominence propagates ethical concerns. The data collectors have an ethical responsibility to ensure datasets curated for machine learning reduce the negative aspects of artificial intelligence.[36] The collectors should ethically capture data and remove unethical components and biases.

Transparency and Explainability

A final note on AI ethics is defined as model transparency or explainability. These terms are interchangeable. There exists a professional obligation to explain how and why a model is making its predictions, especially when the predictions directly impact humans.[37] However, some models are inherently more difficult to understand. Neural networks are notoriously challenging to interpret.[38] Their absent transparency is a direct consequence of their incredibly large number of nodes, connections, and trained weights. The number of parameters can exceed tens of millions in disciplines such as computer vision.[39] The pertinent parameters involved in a neural network's prediction are hard to precisely identify. This is starkly contrasted with different algorithms, like COMPAS, which has at most 137 parameters and was measured to be only slightly superior to a model with two parameters.[40]

There are numerous guidelines on the use of artificial intelligence predictions. Yu Zhang et al. reference an inherent "right to explanation" for predictions as well as the European Union's General Data Protection Regulation, Article 22, for its stated importance of protecting data.[41] Zhang et al. mention an important, clear example for transparency: explaining a medical diagnosis using AI. People receiving medical diagnoses from an AI deserve clear explanations for those decisions.

Transparency is a professional requirement in artificial intelligence. However, AI transparency might be no deeper than what already exists for human decision-making. A doctor will make his or her best diagnosis based upon a thorough medical exam, lab work, imagery scans, and then a comparison to his or her aggregate of experience and training. Medical AI should be expected to do no less. Complex neural networks may indeed lack explainability but then again so do many human decisions. Humans are fond of using expressions like "I had a gut feeling," "My intuition guided me," "It felt just right," or "I knew it in my bones." The lack of explainability does not necessarily correlate with correct or incorrect decision-making, but it is an "alert" for possible biased conclusions for both humans and AI. However complex, AI should be accompanied by fully transparent documentation.[42]

Conclusion

Dataset ubiquity commands ethical consideration. Datasets are the center of gravity in AI ethics, and equitable dataset engineering should be an explicit measure of ethics. Professionally developed datasets are required for artificial intelligence to function properly and ethically. They must be robust and complete. Every demographic in an AI project, from topics as sensitive as race to topic as tactically prominent as target recognition, should receive an equitable treatment. Special care must be taken to ensure an ethical outcome, where the steps in achieving an ethical dataset are relative to use case. ■

This article is an opinion piece of the authors built from their academic and professional experience. The opinions in this article do not reflect that of their units or their professional work.

Notes

1. Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach* (Hoboken, NJ: Pearson Education, 2001), 1006.

2. Geoff Brumfiel, "Israel Is Using an AI System to Find Targets in Gaza. Experts Say It's Just the Start," NPR, 14 December 2023, https://www.npr.org/2023/12/14/1218643254/israel-is-using-an-ai-system-to-find-targets-in-gaza-experts-say-its-just-the-st; Margarita Konaev, *Tomorrow's Technology in Today's War: The Use of AI and Autonomous Technologies in the War in Ukraine and Implications for Strategic Stability* (Arlington, VA: Center for Naval Analyses, 2 October 2023), 7, https://www.cna.org/reports/2023/10/ai-and-autonomous-technologies-in-the-war-in-ukraine.

3. Jack Watling and Nick Reynolds, *Meatgrinder: Russian Tactics in the Second Year of Its Invasion of Ukraine* (London: Royal Uniformed Services Institute for Defence and Security Studies, 19 May 2023), https://www.rusi.org/explore-our-research/publications/special-resources/meatgrinder-russian-tactics-second-year-its-invasion-ukraine.

4. Brumfiel, "Israel Is Using an AI System to Find Targets in Gaza"; Gaza Task Force, *Gaza Conflict 2021 Assessment: Observations and Lessons* (Washington, DC: Jewish Institute for National Security of America, 28 October 2021), https://jinsa.org/jinsa_report/gaza-conflict-2021-assessment-observations-and-lessons/.

5. Ibid.

6. "What Are Edge Cases?," d(risk), accessed 28 May 2024, https://drisk.ai/what-are-edge-cases/; Ofir Zuk, "Edge Cases in Autonomous Vehicle Production," Datagen, 13 April 2022, https://datagen.tech/blog/how-synthetic-data-addresses-edge-cases-in-production/; Lex Fridman et al., "MIT Advanced Vehicle Technology Study: Large-Scale Naturalistic Driving Study of Driver Behavior and Interaction With Automation," *IEEE Access* 7 (2019): 102021–38, https://doi.org/10.1109/ACCESS.2019.2926040.

7. Fridman et al., "MIT Advanced Vehicle Technology Study."

8. Sorin Grigorescu et al., "A Survey of Deep Learning Techniques for Autonomous Driving," arXiv, 24 March 2020, https://arxiv.org/pdf/1910.07738.pdf.

9. Konaev, "Tomorrow's Technology in Today's War," 10–13.

10. Air Force Doctrine Publication 3-60, *Targeting* (Washington, DC: U.S. Government Publishing Office, 12 November 2021), 70.

11. Russell and Norvig, *Artificial Intelligence*, 990.

12. Ibid., 987.

13. "IEEE Strategic Plan 2020–2025," Institute for Electrical and Electronics Engineers (IEEE), accessed 28 May 2024, https://www.ieee.org/about/ieee-strategic-plan.html.

14. Alan F. T. Winfield et al. "IEEE P7001: A Proposed Standard on Transparency," *Frontiers in Robotics and AI* 8 (26 July 2021): Article 665729, https://doi.org/10.3389/frobt.2021.665729.

15. Ibid.

16. Russell and Norvig, *Artificial Intelligence*, 653–720, 881–924. At this point the reader might grow curious as to how a model would or would not detect objects in an image because of training data. The chapters provided in this endnote provide explanations as to how a model learns through data as well as specifics for computer vision. These explanations are somewhat densefor an introduction into the field, but they are comprehensive. These chapters provide more than adequate explanations for the purposes of this article.

17. Rahul Awati, "What Is Garbage In, Garbage Out (GIGO)?," TechTarget, accessed 29 May 2024, https://www.techtarget.com/searchsoftwarequality/definition/garbage-in-garbage-out.

18. Valeriia Cherepanova et al., "A Deep Dive into Dataset Imbalance and Bias in Face Identification," arXiv, 15 March 2022, https://arxiv.org/pdf/2203.08235.pdf/.

19. Russell and Norvig, *Artificial Intelligence*, 992.

20. Ibid, 993.

21. Ethem Alpaydin, *Machine Learning*, rev. and updated ed. (Cambridge, MA: MIT Press, 2021), 50–51.

22. "Harop Loitering Munitions UCV System," Airforce Technology, 2 July 2015, https://www.airforce-technology.com/projects/haroploiteringmuniti/?cf-view; Russell and Norvig, *Artificial Intelligence*, 988.

23. Russell and Norvig, *Artificial Intelligence*, 988–95; Joy Buolamwini and Timnit Gebru, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (PMLR) 81, (2018): 77–91, http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf.

24. Cherepanova et al., "A Deep Dive into Dataset Imbalance and Bias in Face Identification"; Judea Pearl, *Understanding Simpson's Paradox*, Technical Report R-414 (Los Angeles: University of California, Los Angeles, December 2013).

25. Julia Dressel and Hany Farid, "The Accuracy, Fairness, and Limits of Predicting Recidivism," *Science Advances* 4, no. 1 (17 January 2018), https://doi.org/10.1126/sciadv.aao5580; Russell and Norvig, *Artificial Intelligence*, 993–95.

26. Jeff Larson et al., "How We Analyzed the COPAS Recidivism Algorithm," ProPublica, 23 May 2016, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

27. Russell and Norvig, *Artificial Intelligence*, 994; Kelly Walsh et al., "Estimating the Prevalence of Wrongful Convictions" (Washington, DC: National Criminal Justice Reference Service, September 2017), https://nij.ojp.gov/library/publications/estimating-prevalence-wrongful-convictions.

28. Russell and Norvig, *Artificial Intelligence*, 995.

29. Ibid., 1006.

30. Ibid., 993.

31. Cherepanova et al., "A Deep Dive into Dataset Imbalance and Bias in Face Identification."

32. Russell and Norvig, *Artificial Intelligence*, 1015–17.

33. Ibid.

34. Ibid.

35. Alpaydin, *Machine Learning*, 11–12.

36. Russell and Norvig, *Artificial Intelligence*, 1006.

37. Ibid., 997; IEEE, "IEEE Strategic Plan: 2020-2025."

38. Yu Zhang et al., "A Survey on Neural Network Interpretability," arXiv, 5 July 2021, https://arxiv.org/pdf/2012.14261.pdf; Sandareka Wickramanayake, Wynna Hsu, and Mong Li Lee "Towards Fully Interpretable Deep Neural Networks: Are We There Yet?," arXiv, 24 June 2021, https://arxiv.org/pdf/2106.13164.pdf.

39. Kaiming He et al., "Deep Residua Learning for Image Recognition," arXiv, 10 December 2015, https://arxiv.org/pdf/1512.03385.pdf.

40. Dressel and Farid, "The Accuracy, Fairness, and Limits of Predicting Recidivism."

41. B. Goodman and S. Flaxman, "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation,'" *AI Magazine* 8, no. 3 (2017), https://doi.org/10.1609/aimag.v38i3.2741; Zhang et al., "A Survey on Neural Network Interpretability."

42. Winfield et al., "IEEE P7001."

US ISSN 0026-4148