Aryan Patel

Mrs. Russell

Independent Study and Mentorship

11 September 2026

Annotated Bibliography on Ethics and Deployment Risk in Artificial Intelligence

Key

Introduction, yellow

Aims and Research Methods, green

Scope, blue

Usefulness, purple

Limitations, red

Conclusions, orange

Reflection, teal

Buolamwini, Joy, and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, vol. 81, 2018, pp. 77-91. *Proceedings of Machine Learning Research*, proceedings.mlr.press/v81/buolamwini18a.html.

Introduction. Buolamwini and Gebru used their own face dataset to test three commercial face analysis products from Microsoft, IBM, and Face++. Rather than providing one accuracy figure for each product, they provided the accuracy figures by skin tone and gender label. The overall score was good for all products, and all products suffered from a poor score on one group.

Aims and Research Methods. The authors wished to demonstrate that there is more to a single accuracy number. They put together the Pilot Parliaments Benchmark, which is 1,270 official photos of national parliament members, and labeled each face with a Fitzpatrick skin type. The labels were checked by a dermatologist. They then used each company's public API to run all the photos through and reported error rates for four groups rather than one consolidated score.

Scope. The data is for six countries, three in Africa and three in Europe. These were the countries that they selected because they wanted a large enough number of darker skin faces to be able to measure anything. The factors involved are skin type, gender label, and which company produced the answer. They chose parliament photos because they are public, they were framed the same way and already have a name and a country attached.

Usefulness. The number that I want to keep going back to is the gap. The lowest performing product missed 0.8 percent of the time on lighter skinned men and 34.7 percent of the time on darker skinned women. I write vision code for my robotics team and I have never divided my test results by anything. I just see the accuracy and then if it went over the target, I move on. This paper made me realize that my testing would allow that same failure to get through without warning me.

Limitations. The data set is limited and the images are clean official headshots with even lighting. The paper's error rates are likely to be lower than real world error rates in an airport or police database. The authors acknowledge that gender is presented as a dichotomy, and this restricts the meaning that the numbers can have. The results are also from 2017 and all three companies have updated their systems since the paper was published, so the percentages are not necessarily current, even though the method still is.

Conclusions. I am making a change in my testing of vision models I write. I want a few slices for each test set, starting with lighting and distance, which is what my robot has to deal with. This also gave me a research direction I did not have before. When I joined ISM I knew I wanted to study machine learning and nothing narrower than that. Now my question is how you catch a model that fails for one group of people before you ship it.

Reflection. I chose Artificial Intelligence for ISM because I enjoy creating something that functions. This paper is about what occurs when no one says who the system is supposed to work for. That question sits under every project I've created, including the ones that I am proud of, and I would prefer to learn to ask it now rather than after I've written something that's running somewhere that matters.

Mitchell, Margaret, et al. "Model Cards for Model Reporting." *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, 2019, pp. 220-29. *ACM Digital Library*, doi.org/10.1145/3287560.3287596.

Introduction. Mitchell and her coauthors suggest a short document, one or two pages long, that is shipped with a trained machine learning model. The model card details the purpose of the model, the population it was tested with, how it performed on each population, and what the authors believe it should not be applied to. This isn't so much a discovery as a suggested habit.

Aims and Research Methods. The goal is to make the model's training domain and the actual domain as similar as possible. The authors work backward and start by identifying the failures that occur because of that gap, and then create a template to identify those failures. The template includes set fields for model details, intended use, factors, metrics, evaluation data,

training data, ethical considerations and caveats. They then complete it twice, once for a face detection model and once for a text toxicity classifier, allowing readers to see a completed card rather than debate an idea.

Scope. Whether it's a public API or handing it over to another team in the same company, the proposal is for any trained model that is released. The real work is done in the factors field. It requires the author to identify groups and conditions that could affect performance, and report a number for each group instead of a pooled score.

Usefulness. This is the most directly usable source that I found. The other papers tell me that there is a problem. This one hands me a form. I sat down and filled out the fields for the object detection model I want to use for my robotics team and got stuck right at the intended use section, because I have never thought about what conditions the model is supposed to handle. Getting stuck was the useful bit.

Limitations. Nothing makes anyone write a model card, and nothing stops someone from writing a vague card. The paper acknowledges this and it is not addressed. A card also only reports on the groups the author was expecting to find failures in, and says nothing about the groups the author did not think to check. That is a beginning, not an end.

Conclusions. I want to make an actual model card for my vision model before the end of this semester, and then use it as an artifact for ISM. That will force me to make decisions that I've been putting off, such as what the acceptable miss rate is. It also brings me back to Buolamwini and Gebru, where the factors field is the same place where their group by group evaluation would reside.

Reflection. I have been treating documentation as a product of the successful project. This paper presents the alternative view, namely that it is a part of the modelling process to write

down the intended use and the known limits. I believe that's correct. If I continue in this line of work, writing the card first is probably more valuable to me than any one technique I learn.

National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce, Jan. 2023. NIST AI 100-1, nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf.

Introduction. This is a government document, not a research paper. NIST outlines a voluntary approach to identify and mitigate risks in AI systems that organizations develop or acquire. It was like a long list with a lengthy explanation attached, which was different than all other reading I had done.

Aims and Research Methods. It's designed to provide organizations with a common language and a process that they can repeat rather than have each company create its own. There are two parts to the framework. The first is the definition of risk and the things NIST believes a trustworthy AI system should have, such as validity, safety, security, accountability, transparency, explainability, privacy, and managed bias. The second part breaks this down into four functions, which are govern, map, measure and manage. The document was developed using an open comment process, rather than being developed internally by NIST and published as complete.

Scope. The framework covers every sector and the whole life of a system, from design through deployment and retirement. It avoids specifying a particular technique or numeric threshold, since a rule that applies to a medical device wouldn't apply to a recommendation feed. This is the only one I've read that is not about the model, but the organization.

Usefulness. This one was not as exciting to me as I'd hoped, but it was more useful. It is dry. It did answer a question that the research papers left unanswered, however, and that was who

is supposed to do something about a biased model. The govern function makes it explicit and specifies it in terms of people and policies rather than who happens to notice. The map function also gave me language for something I have been doing wrong, which is setting the context before I start building.

Limitations. The framework is voluntary and does not specify anything, so a company can claim to follow the AI RMF without taking any action. Throughout the document, there is no threshold, no audit and no consequences. Also, it makes the assumption that it's an organization with staff and process, not a two person project or a high school robotics team, so most of it had to be translated before I could apply it to myself.

Conclusions. I'm going to take the four functions as a guideline for my own project this year, primarily to get me to commit the desired context to paper before I start coding. I'm also interested in how this is different from the EU AI Act, which is mandatory, and that could be another study to be done later.

Reflection. Just about everything I know about AI came from tutorials and papers, so I know how systems get built and almost nothing about how they get approved. This is the first thing I've read from the approval side. If I find myself working on something that's being deployed, the people who are asking me questions will be talking in this way, and I'd rather be familiar with it than have to learn it in the room.

Obermeyer, Ziad, et al. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science*, vol. 366, no. 6464, 25 Oct. 2019, pp. 447-53. doi.org/10.1126/science.aax2342.

Introduction. Obermeyer and his coauthors looked at a commercial algorithm that hospitals use to decide which patients get extra care management. The algorithm was already in use for millions of people. The authors were able to obtain the underlying data from a large academic hospital, and they discovered that the system was overlooking Black patients who were sicker than the white patients it was flagging instead.

Aims and Research Methods. The authors wished to demonstrate the mechanism, rather than simply that there was a gap. They analyzed records from about 50,000 primary care patients, including their risk scores, insurance claims, and clinical data such as their blood pressure, lab tests, and other metrics. They could ask themselves if those patients were really equally sick when they compared patients who received identical risk scores. They were not. It was actually the target variable that caused the issue. The model was used to predict future health care expenditures, and the hospital used that prediction as an indicator of future health care needs.

Scope. The study is limited to the patients of one health system over a number of years. The primary variables are race, risk score, claims cost and measured illness. That makes this more than a complaint, as the authors then simulated what would occur if the model predicted health conditions rather than costs.

Usefulness. It was the mechanism that got me. No one included race in the model. The inputs were neutral, the engineers were not negligent, and the model did very accurately what it was trained to do. The bias was introduced by using cost as the label, because less money is

spent on Black patients at the same level of illness, so the model learned that they required less care. The authors' simulation showed that the percentage of Black patients identified for additional care went from approximately 18 percent to approximately 47 percent when the target was changed. That's a huge swing out of a decision most teams would make in a meeting without discussing it.

Limitations. The analysis is based on one hospital system, so the numbers may not apply to other systems, but the authors say that the same design decision is prevalent throughout the industry. The vendor is never identified, which makes accountability hard, even though staying unnamed is probably what made the research possible. The paper also uses claims and lab measures of illness, which are similarly distorted by who has access to care, so even after the correction there is some distortion.

Conclusions. I want to focus on the label and not the architecture anymore. If I write down a problem I'll write down the target variable and the thing I actually care about as two separate lines and see if they match. This whole failure lived between those two lines, and I would never have thought to look there.

Reflection. This is the source that changed my mind about where AI risk comes from. I was thinking of some bad data or careless engineers. This was neither. It was a reasonable substitute that capable people selected and then forgot about, which is a mistake I'm very capable of making. That's why it was more memorable than the papers that described more dire systems.

Rudin, Cynthia. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence*, vol. 1, no. 5, May 2019, pp. 206-15. doi.org/10.1038/s42256-019-0048-x.

Introduction. Rudin claims that the field veered off course. If a very complicated model makes a decision that has an impact on somebody's life, the usual reaction is to create a second model that explains the first one. She says this is the wrong way to go about it, and that models that people can understand should be constructed first in the case of high stakes decisions.

Aims and Research Methods. This is an argument paper, not an experiment, but she constructs her argument from evidence, not assertion. She distinguishes between explaining a black box and designing a model that is interpretable, and makes her central claim. A post hoc explanation is an approximation of the actual model, so it is wrong some of the time, and an explanation that is wrong some of the time is worse than useless in a courtroom or a hospital. To support this, she points to criminal risk assessment, where her own team developed simple models based on plain rules which performed at a level comparable to proprietary black box models on the same data.

Scope. She focuses on high stakes decisions such as criminal sentencing, loan approval, medical treatment and parole. She says straight out that her argument doesn't apply to low stakes uses such as ad ranking, where accuracy matters more than someone's ability to argue with the result. She set that limit on purpose, so it is part of the argument.

Usefulness. This one disagreed with a belief of mine. I thought that accuracy and interpretability traded off against each other and that using a black box would have to be a compromise to get good performance. But according to Rudin, that trade off often doesn't exist on structured data with features that have meaning, and she has evidence to support her claim. I

have not tried it myself and I am not convinced, but I can't rule it out either, and I stopped repeating the trade off as if it were a done deal.

Limitations. Her evidence is largely tabular data with features someone designed carefully, which is where simple models do their best work. The argument is more difficult to make with images or text, and she doesn't really make it there. What is considered high stakes is also left to the reader's discretion, which is nice for those who feel they can call their own product low stakes and keep the black box.

Conclusions. I don't want to take her word for it, so I want to test it on something of my own. If I develop a classifier this year, I want to start with a simple interpretable one and if it's obviously losing, I will go to something heavier. That's a real answer, not an opinion, and something I can show a mentor.

Reflection. I like this paper because it is contrary to the direction most of the papers I read go. Nearly everything I've taught myself has pushed towards larger models, and here's a serious researcher saying that instinct is a hindrance in the cases that matter. Learning to sit with a source that is contrary to me is probably worth as much to my ISM year as the argument itself.