Research Statement

Jamelle Watson-Daniels [*]
Harvard University

Research Agenda: My research vision is to develop principled approaches for measuring and evaluating algorithmic decisions made at each stage of the machine learning pipeline. I want to contribute to a collaborative interdisciplinary research program where we bridge the growing gap between technical advances and socio-political conditions. Looking ahead, I plan to be at the forefront of developing methods that promote fairness and reliability in generative Ai.

From initial conception to deployment, researchers and practitioners make critical decisions at each stage of the machine learning pipeline, the significance of which can be quantified by evaluating alternative options. For instance, there can be multiple models with near-optimal performance for a given prediction task. If model outputs vary significantly between these similar models, then the decision to choose one model over another has relative importance. Further, the model selection decision might come under scrutiny if there exists a similar model with better fairness properties that could have been selected. Even long before deployment, when translating high-level goals into tractable predictive tasks, there may be many reasonable target variable options worth considering. Once again, target specification becomes particularly high stakes when one target leads to more disparate treatment compared to other options. In each case, the significance of these decisions can be characterized by understanding what changes over possible alternatives. In what follows, I outline my previous research and a few future directions.

1 Predictive Multiplicity

In machine learning, *model multiplicity* is the existence of multiple models that perform equally well for a given prediction task (also known as the "Rashomon effect"). This set of near-optimal models with similar performance but different characteristics is referred to as the "Rashomon set". *Predictive multiplicity* examines how predictions change over this Rashomon set. My previous research introduces frameworks for studying predictive multiplicity in different settings.

As in the standard predictive multiplicity setting [11], I begin with a *baseline model*, h_0, that is the solution to an empirical risk minimization (ERM) problem of the form $\min_{h \in \mathcal{H}} L(h; \mathcal{D})$, over a hypothesis class, $\mathcal{H}$, with loss $L(\ \cdot\ ; \mathcal{D})$. In this context, one can consider the ϵ-Rashomon set, which is the set of all models that achieve near-optimal loss.

For a baseline model h_0 and error tolerance $\epsilon > 0$, the ϵ-Rashomon set of competing models is the set of classifiers that satisfy $L(h; \mathcal{D}) \leq L(h_0; \mathcal{D}) + \epsilon$. In [11], $\mathcal{H}$ is assumed to be a class of binary classifiers and one of the predictive multiplicity measures introduced *ambiguity* of a prediction problem: the proportion of points in the training dataset assigned conflicts ($\mathbb{1}[h(x_i) \neq h_0(x_i)]$) over the ϵ-Rashomon set of competing models. The *discrepancy* of a prediction problem is the maximum proportion of points in the training dataset assigned conflicts ($\mathbb{1}[h(x_i) \neq h_0(x_i)]$) by the single worst case competing model.

Predictive Multiplicity in Probabilistic Classification [21]: Probabilistic classification is often incorporated into real-world risk assessment tasks to inform decisions. For instance, probabilistic classifiers that predict consumer default risk are used by lenders to underwrite loans. I developed a framework for investigating predictive multiplicity in this setting. Measuring multiplicity in probabilistic classification is complicated by the need to clarify the meaning of "conflicting". In effect, what constitutes a conflicting risk prediction can change across applications (e.g., predictions that vary by 5% or 30%). Likewise, what constitutes a "competing" model can change across applications. My work addresses both of these problems by introducing methods that allow users to specify what is "competing" (near-optimal metric) and what is "conflicting" (deviation threshold).

[*] Harvard University. email: jwatsondaniels@g.harvard.edu

To this end, I consider loss, accuracy and calibration error as possible near-optimal metrics and redefine ambiguity and discrepancy in this setting. I also introduce the *viable prediction range* which captures how individual predictions change over the Rashomon set.

The *viable prediction range* is the smallest and largest risk estimate assigned to example i over competing models in the ϵ-level set. The (ϵ, δ)-*ambiguity* of a probabilistic classification task over a sample S is the proportion of examples in S whose baseline risk estimate changes by at least δ over the ϵ-level set. The (ϵ, δ)-discrepancy of a probabilistic classification task over a sample S is the maximum proportion of examples in S whose risk estimates could change by at least δ by switching the baseline model with a competing model in the ϵ-level set.

Methodology: Our optimization based methods compute our measures reliably. To compute ambiguity and viable prediction ranges, I construct a pool of *candidate models* that assign a specific risk estimate to each example. From these models, I select those that have performance within ϵ of the baseline model as the set of competing models. Specifically, for each threshold probability $p \in P$, I train a candidate model h such that the probability assigned to the example is constrained to the threshold p. To compute discrepancy, I formulate a mixed-integer non-linear program (MINLP) which involves constructing a linear approximation of the loss using an iterative, outer-approximation method to solve. This method is exact for computing discrepancy in terms of near-optimal loss. For other metrics, we can again treat the intermediate solutions to the outer-approximation algorithm as candidate models and use these candidates to recover a lower bound similar to the method used to compute ambiguity and viable prediction ranges. Using synthetic data, I also presented the first study providing insight into the kinds of data characteristics that give rise to predictive multiplicity.

Predictive Multiplicity Under Resource Constraints [17]: In this work, I introduce a framework for assessing predictive multiplicity in the presence of decisions under resource constraints to extend previous analysis to predictive allocation tasks. In practice, there is often only a finite amount of benefit, burden, or scrutiny that a system is able to allocate. This means that the set of "good models" should only include models that satisfy the resource constraint. In this work, I define a new measure of predictive multiplicity (top-κ ambiguity) and present a mixed integer program (MIP) to calculate this ambiguity measure for linear models. Note that in binary classification [11], a prediction problem could have high ambiguity if the positive classification rate, $\frac{1}{n}|\{i : h(x_i) = 1\}|$, differs greatly between h_0 and models in $\mathcal{H}_\epsilon(h_0)$. That is, a high ambiguity may simply result from models that allocate a very different number of resources. Thus, extending to incorporate resource constraints enhances predictive multiplicity research in a number of ways.

Given a prediction model h and resource cap κ, I let $Top_{(i,h,\kappa)} = \mathbb{1}[\tau_i(h) \leq \kappa]$ be the indicator of whether instance i is "in the top-κ" when ranked according to the predicted values h. This work defines two notions of ambiguity in this setting. The (ϵ, κ)-*ambiguity (all)* over a sample, S, is the proportion of examples for which the top-κ decision changes over the ϵ-Rashomon set. The (ϵ, κ)-*ambiguity (top)* over a sample, S, is the proportion of top-κ examples according to h_0 that fall outside the top-κ for some models in the ϵ-Rashomon set.

Methodology Enhanced: My prior work involves constructing a pool of candidate models that change individual predictions [21]. From that pool of models, those with near-optimal performance are selected to compute ambiguity. These methods are indirect in that the MIPs do not directly constrain these candidate models to be within the ϵ-Rashomon set. Under resource constraints, I develop a MIP that **does** include this constraint by theoretically showing how to include a constraint that neatly characterizes the ϵ-Rashomon set for linear models. Additionally, I show theoretically that (i) one can efficiently determine that many points are provably *not flippable* over the ϵ-Rashomon set; and (ii) one can identify a subset of *flippable* points by solving a proxy optimization problem with a closed-form solution that produces a $w \in \mathcal{H}_\epsilon(w_0)$ that *may* flip some points into the top-κ. This means that in practice, we only need to solve the computationally expensive MIP for a very small subset of points whose flippability remains undetermined following the efficient filtering steps.

2 Target Specification

Multi-Target Multiplicity: Flexibility and Fairness in Target Specification under Resource Constraints [17]: Real-world problems rarely present themselves as fully formed machine learning tasks [13]. Critically, it is often not clear what target should be predicted to help decision makers achieve their goals [7, 12]. It is far from obvious, for example, how employers should go about making such choices in their hiring practices: if the goal is to hire the "best" people, what exactly should the model be predicting [1, 9, 12]? For a sales position, employers might choose to predict annual sales figures. But they could alternatively choose to predict how well the applicant will work with others, whether customers will actually enjoy interacting with the applicant, etc. Even in domains where target choice might seem more obvious, there can still be a good deal of uncertainty. For example, while it might seem self-evident that creditors should be predicting default, what constitutes "default" is not a given. Creditors need to make an affirmative choice about the number of months of missed payments that ultimately count as "default" [8]. In some cases, the decision is not based on just one chosen target, but instead a combination of targets. For example, many algorithmic tools currently used in criminal justice and human services function by aggregating predictions of several different targets, ranging from different types of criminal justice system encounters, to mental and physical health outcomes, to measures of housing stability [10, 14].

A recent line of work has explored the implications of this flexibility in target variable choice for fairness. Prior work does not, however, offer a more general mathematical or computational framework for characterizing the extent to which target variable choice affects individuals' outcomes and selection rate disparities across groups. My work fills this gap [17]. By analogy to predictive multiplicity, in the motivating "multi-target" setting—where there are many possible reasonable prediction targets to choose from—we can consider the set of "good models" that arises from, say, models that predict *any* of the individual targets well, models that predict a *combination* of targets well, or that *combine the predictions* of single-target models. If each of the possible targets individually has merit as a basis for predictive allocation, then any *combination* of those predictions is arguably also a reasonable candidate model.

3 Racism and Algorithms

The realization that algorithms can perpetuate or exacerbate racial disparities in society has spurred significant research in the field of algorithmic fairness. While contending with racism has been a primary motivation and driver of research in this area, developing strategies to correct and prevent racist outcomes is an ongoing challenge partially because modern forms of racism have evolved to be quite nuanced. Even outside of algorithms, modern racism tends to be concealed by seemingly "race-neutral" methods and rhetoric making it hard to identify and therefore address. How do we solve a problem that we cannot see? Scholars often refer to this modern form of racism as *colorblind racism* which occurs when we observe a racially discriminatory outcome but the mechanism responsible for said outcome appears to have nothing to do with race [3].

This colorblind racism becomes particularly enhanced when coupled with modern advances in technology [2]. Similar to work on colorblind racism, there is an emphasis in recent studies on clarifying the concealed nature of racism functioning in the context of algorithms. While this type of work tends to span various technological disciplines i.e. robotics, search algorithms, etc., there is limited work on how this more subtle form of racism arises within specific sub-fields of computer science. I have been exploring anti-Black racism primarily to facilitate understanding and clarify terms of discussion on the topic in the algorithmic setting. In my recent work, I discuss colorblind racism in the context of algorithmic fairness research [16].

Since my dual undergraduate concentration in Africana studies and Physics, I have maintained a commitment to collaborations across disciplinary lines. Following the 2020 Black Lives Matter protests, my collaborators and I used a few-shot domain adaptation approach from natural language processing to reveal the prominence of positive emotions (encompassing, e.g., pride, hope, and optimism) in tweets with explicit pro-BlackLivesMatter hashtags and correlated with on the ground protests [4]. This work illustrates the power of online activism in support of social movements. Joining a collective call to support Black public health experts and community organizers, I helped organize and document community conversations on using data as a tool for social change during the 2020 pandemic [20]. Also, I led a team of software engineers and data scientists in a public service project where we documented the challenges in automated web-scraping of COVID-19 data from US state websites [19]. Another non-profit collaboration, I contributed to an examination of *data capitalism* to help policymakers and activists understand how the extraction and commodification of data as fundamentally intertwined with systemic racism [5]. I look forward to future opportunities to contribute to these interdisciplinary collaborations.

4 Future Directions

Predictive Multiplicity in Generative Ai, Recommender Systems, Graphs, etc: I plan to study predictive multiplicity in the generative Ai setting, which requires a slightly different framing of the problem. Following the existing definition, I also plan to develop a framework for analyzing predictive multiplicity for recommender systems and graph based models. I am generally interested in exploring additional settings where model multiplicity invokes interesting methodological opportunities.

Predictive Churn: ML models in modern applications are often updated over time. One of the foremost challenges faced is that, despite increasing overall performance, these updates may flip specific model predictions in unpredictable ways. In practice, researchers quantify the number of unstable predictions between models pre and post update – i.e., *predictive churn.* In recent work [18], I study this effect through the lens of *predictive multiplicity* – i.e., the prevalence of conflicting predictions over the set of near-optimal models (the ϵ-Rashomon set). I show how traditional measures of predictive multiplicity can be used to examine expected churn over this set of prospective models – i.e., the set of models that may be used to replace a baseline model in deployment. Further, I show that our approach is useful even for models enhanced with uncertainty awareness. In the future, I plan to investigate how churn reduction methods might lessen predictive multiplicity.

Reliable Deep Learning: Predictive multiplicity can be conceptualized as a type of predictive arbitrariness as unfairness. Beyond the lens of fairness, this predictive inconsistency or arbitrariness is also a practical concern in the deployment of many production machine learning models. What is the relationship between uncertainty quantification methods and arbitrariness as unfairness (i.e. predictive multiplicity)? Will methods from reliable deep learning prove to be more or less robust to predictive multiplicity? Can uncertainty-aware models provide a signal for potential predictive instability downstream? I am actively investigating these questions and plan to continue to do so.

Fairness in Generative Settings: Recent advancements in large generative models are influencing algorithmic fairness discussions and current methods of fair classification are not immediately relevant to the generative setting. I plan to develop methods that promote fairness and reliability in generative Ai. In particular, I want to explore policy and downstream implications of red-teaming methods which are adversarial probing techniques used to induce harmful outputs then for updating the model to circumvent the harmful outputs. Initial studies [6] examine the relationship between model size, model type and attack success. Building on this type of work, I want to dive deeper into methods for ensuring safety in this setting. There are also interesting interpretability and reliability questions in the context of multimodal learning (the combination of different data modalities i.e. incorporating both image and text input). With collaborators at Georgia Tech, I am studying cross-modal projections for fine-tuning domain-specific models [15].

References

[1] Solon Barocas and Andrew D Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104 (2016), 671.

[2] Ruha Benjamin. 2019. Race After Technology: Abolitionist Tools for the New Jim Code. *Social Forces* 98, 4 (12 2019), 1–3. https://doi.org/10.1093/sf/soz162 arXiv:https://academic.oup.com/sf/article-pdf/98/4/1/33382045/soz162.pdf

[3] E. Bonilla-Silva. 2013. *Racism without Racists: Color-Blind Racism and the Persistence of Racial Inequality in America*. Rowman & Littlefield Publishers. https://books.google.com/books?id=jPBdz1ykpagC

[4] Anjalie Field, Chan Young Park, Antonio Theophilo, Jamelle Watson-Daniels, and Yulia Tsvetkov. 2022. An analysis of emotions and the prominence of positivity in #BlackLivesMatter tweets. *Proceedings of the National Academy of Sciences* 119, 35 (2022), e2205767119. https://doi.org/10.1073/pnas.2205767119 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.2205767119

[5] Data for Black Lives and Demos. 2021. Data Capitalism and Algorithmic Racism. https://www.demos.org/research/data-capitalism-and-algorithmic-racism

[6] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. arXiv:2209.07858 [cs.CL]

[7] David J Hand. 1994. Deconstructing statistical questions. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 157, 3 (1994), 317–338.

[8] David J. Hand. 2006. Classifier Technology and the Illusion of Progress. *Statist. Sci.* 21, 1 (2006), 1 – 14. https://doi.org/10.1214/088342306000000060

[9] Pauline T Kim. 2022. Race-Aware Algorithms: Fairness, Nondiscrimination and Affirmative Action. *California Law Review* 110 (2022), 1539.

[10] Chamari I Kithulgoda, Rhema Vaithianathan, and Dennis P Culhane. 2022. Predictive risk modeling to identify homeless clients at risk for prioritizing services using routinely collected data. *Journal of Technology in Human Services* 40, 2 (2022), 134–156.

[11] Charles Marx, Flavio P. Calmon, and Berk Ustun. 2019. Predictive multiplicity in classification.

[12] Samir Passi and Solon Barocas. 2019. Problem formulation and fairness. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* (2019), 39–48. https://doi.org/10.1145/3287560.3287567

[13] Foster Provost and Tom Fawcett. 2013. *Data Science for Business: What you need to know about data mining and data-analytic thinking.* O'Reilly Media, Inc.

[14] Arnold Ventures. 2022. What is the PSA? https://advancingpretrial.org/psa/about/

[15] Gaurav Verma, Minje Choi, Kartik Sharma, Jamelle Watson-Daniels, Sejoon Oh, and Srijan Kumar. 2024. Mysterious Projections: Multimodal LLMs Gain Domain-Specific Visual Capabilities Without Richer Cross-Modal Projections.

[16] Jamelle Watson-Daniels. 2024. Algorithmic Fairness and Color-blind Racism: Navigating the Intersection. arXiv:2402.07778 [cs.CY]

[17] Jamelle Watson-Daniels, Solon Barocas, Jake M. Hofman, and Alexandra Chouldechova. 2023. Multi-Target Multiplicity: Flexibility and Fairness in Target Specification under Resource Constraints. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 297–311. `https://doi.org/10.1145/3593013.3593998`

[18] Jamelle Watson-Daniels, Flavio du Pin Calmon, Alexander D'Amour, Carol Long, David C. Parkes, and Berk Ustun. 2024. Predictive Churn with the Set of Good Models. arXiv:2402.07745 [cs.LG]

[19] Jamelle Watson-Daniels and Data for Black Lives. 2020. The Impact of COVID-19 on Black Communities. `https://d4bl.org/datasets/6-the-impact-of-covid-19-on-black-communities`

[20] Jamelle Watson-Daniels, Yeshimabeit Milner, Nicole Triplett, Irene Headen, Dominique Day, Zinzi Bailey, Meme Styles, Lisa Clinton, Courtni Andrews, Michelle Wilson, Nchedochukwu Ezeokoli, Stacy Jebbett Bullard, and Lucas Mason-Brown. 2020. Data for Black Lives COVID-19 Movement Pulse Check and Roundtable Report. `https://d4bl.org/reports/41-data-for-black-lives-covid-19-movement-pulse-check-and-roundtable-report`

[21] Jamelle Watson-Daniels, David C. Parkes, and Berk Ustun. 2023. Predictive Multiplicity in Probabilistic Classification. *Association for Advancement in Artificial Intelligence (AAAI)* (2023). `http://arxiv.org/abs/2206.01131`