FORMALIZING PREDICTION RESEARCHERS' CONSIDERATION OF INTERVENTIONS

Kartikeya Kandula and Rushi Shah

ABSTRACT

This essay provides a critique of the prediction research community's failure to sufficiently grapple with prediction based interventions, and a suggestion for a partial remedy. We begin by motivating our criticism and introducing a set of demonstrative case studies from predictive research applied to four problem domains (genetic defect screening, criminal risk assessment, healthcare diagnoses, and family life outcomes). We then propose a worksheet for prediction researchers to use in the design phase of their research to reflect on how they can maximize the intentional and minimize the unintentional downstream consequences of their research. Using the case studies we outlined, we justify and highlight the motivation for each question we pose. After outlining related work on research criticism, we briefly conclude.

1 Introduction

In 2015, Philip Rogaway issued a call to fundamentally transform the culture of cryptography [25]. In the wake of the Snowden disclosures, he considered it a failure of the field that ordinary people "lack even a modicum of communication privacy when interacting electronically" while cryptographers tinker with puzzles and interesting math problems. Despite the mathematical nature of cryptography, Rogaway claims that researchers in the field have a significant social responsibility because of the inherently political nature of how cryptography is utilized.

Within the field of prediction, we have reached a similar crossroads. Recent advances in machine learning have resulted in many optimistic claims about the potential of predictive models to enable valuable interventions in domains ranging from predicting life outcomes [26] to advertisement click through rates [19]. However, it is clear that predictive abilities currently fall well short of the hype. Even when predictive technologies are successfully implemented in interventions, they can have harmful impacts [2]. Unfortunately, researchers studying predictive technology often fail to grapple with how their work can enable progress and with how their work can be misused.

As researchers continue using predictive modeling in various domains, it is vital that they consider the downstream impact their work may have. With the rest of this paper, we introduce a set of demonstrative case studies from predictive research applied to four problem domains (genetic defect screening, criminal risk assessment, healthcare diagnoses, and family life outcomes). We then propose a worksheet for prediction researchers to use in the design phase of their research to reflect how they can maximize the intentional and minimize the unintentional downstream consequences of their research. After using the case studies to demonstrate the value of our proposed worksheet, we close our paper by discussing related work.

2 Case Studies

We will use the following four problem domains, and the predictive research applied to each, to illustrate the claims we make in the rest of this paper.

2.1 Genetic Defect Screening

"As advances in medical technology and genetic science increases, eugenics is making a return into the American psyche.", according to B. De Neice Welch. Intentionally, or not, researchers creating screenings for predicting genetic defects are legitimizing the concept of neo-eugenics. This screening represents "a soft coercion toward a eugenic bias. Should an abnormality be discovered, there is a presupposition that the pregnancy will be terminated, and some physicians will refuse to perform amniocentesis unless termination is agreed to prior to performing the test." [28]

Such advancing technology led Armand Marie Leroi to speculate that "if a geneticist were able to screen a randomly chosen embryo for all known disease genes [...] the probability of predicting an inherited disease in a randomly chosen human embryo is almost 0.4% (Table 1). Therefore, it should be possible to predict a disease in 1 in 252 embryos."[18] Leroi even acknowledges there are significant downstream consequences of such research, but dismisses the topic with a plea of ignorance. He says "Some readers might find it peculiar that in this discussion of neo-eugenics, I have not considered the ethical or legal implications with which this subject is generally considered to be fraught. Although I do not doubt their importance, I simply have no particular knowledge of them." This cavalier attitude towards the ethical implications of neo-eugenics is representative of many scientists' negligence towards the broader impacts of their work.

2.2 Criminal Risk Assessment

Similarly, risk assessment researchers are legitimizing the criminal legal system they purportedly aim to reform. As Ben Green describes, "by tweaking surface-level decisions and providing them with a semblance of neutrality and fairness, risk assessments are likely to sanitize, legitimize, and perpetuate the criminal justice system's carceral and racist structure. [...] This process of legitimation can be seen most clearly with regard to preventative detention (detaining a criminal defendant before trial due to concerns about crime risk)." [13] In the American legal system, the accused are presumed innocent until proven guilty, but risk assessment technologies provide justification for incarcerating individuals before their trial. Not only does such pretrial incarceration have "far reaching negative consequences" from a policy perspective [11], but it also perpetuates racial inequities in the criminal legal system [2]. Therefore, rather than reforming the criminal legal system, risk assessment research is providing quantitative legitimacy to the system's worst aspects.

2.3 Healthcare Diagnoses

In some cases, the intended intervention from a predictive model can be explicit and worthwhile to pursue and still lead to work that is flawed due to researchers not adhering to standards of the domain they are engaging with. In the evaluation of retinal fundus photographs from adults with diabetes with an algorithm based on deep machine learning, a study led by researchers at Google

[14] reported high sensitivity and specificity for detecting referable diabetic retinopathy. Despite these optimistic results, however, there are many elements of this study that diminish its real world clinical relevance and are emblematic of issues of prediction in the context of healthcare.

In an attempt to replicate this study, Voets et al. [27] had to re-implement the methodology and utilize alternative sources of data because the original study did not provide source code or the data sets that were utilized. The original study used non-public fundus images for training and a separate data set for evaluation that is no longer available, forcing Voets et al. to turn to publicly available substitutes. Because of this deviation from existing reporting standards, the results achieved by the replication study did not come close to the results of the original study. The lack of reproducibility from the original Google study is not merely an isolated instance in artificial intelligence research. In another study led by Google researchers, a deep learning model was reported to outperform human experts in breast cancer prediction from mammograms [20]. Critics of the study, however, assert that "the Google team provided so little information about its code and how it was tested that the study amounted to nothing more than a promotion of proprietary tech." [17] In a response to the study, 31 scientists posted a letter critical of the lack of reproducibility of the Google study and stated that the lack of access to code and data in prominent scientific publications may lead to unwarranted and even potentially harmful clinical trials. [15]

Along with these replication issues, the original study was not prospective and did not occur in a real world clinical environment. In spite of evidence of the superiority of actuarial methods over clinical methods [9], such differences in the diagnostic environment results in comparisons of the machine learning model performance against human clinicians that are difficult to evaluate. Although this study was not meant to prove effectiveness in a clinical setting, this difference in experimental environment is representative of research into diagnostic deep learning algorithms for medical imaging and can lead to exaggerated claims of machine learning performance. This presents a risk to patient safety and population health at the societal level when such algorithms are applied to patients at scale [21].

2.4 Family Life Outcomes

The Fragile Families challenge showed that hundreds of researchers were unable to accurately predict six life outcomes, such as a child's grade point average and whether a family would be evicted from their home, despite drawing on a variety of machine learning methods over a vast data set. One of the strengths of the Fragile Families Challenge paper is how clear it made the conclusions policymakers should draw. Based on the poor predictive accuracy across the board, they instructed policymakers to be hesitant to adopt such predictive techniques. They recommended that before using complex predictive models, "policymakers determine whether the achievable level of predictive accuracy is appropriate for the setting where the predictions will be used, whether complex models are more accurate than simple models or domain experts in their setting (26–28), and whether possible improvement in predictive performance is worth the additional costs to create, test, and understand the more complex model (26)." [26].

Rather than only aiming to create an accurate predictive model, the Fragile Families Challenge also aimed to understand the discipline of prediction itself. In this regard, the research methodology was particularly strong because the common task method addressed this goal well. Also, because of this explicitly-stated larger objective, it is clear to the reader why the study did not provide as

much depth in analyzing potential interventions as it would have if the goal had been to simply predict life outcomes for intervention.

Although we recognize and respect the researcher's intentions, we still wonder what takeaways could be drawn from a similar study that was able to create a predictive model that accurately forecasts, say, a student's future GPA. We aren't sure how this prediction could be used in a realistic intervention. Of course it depends on what features the model uses, how accurate the model is, and what outcome the model successfully predicts. But even with all that information, we wonder what next steps the researchers would desire of policymakers. The Fragile Families Challenge could have addressed this alternative motivation. The authors could have provided their thoughts on what their takeaways from the study would have been if they had instead found strong indicators of predictability.

Although readers such as ourselves may differ from the authors in intended application of the research, the authors commendably set the readers' expectations on what the authors themselves were intending to learn from their work. Furthermore, their research methodology was well tailored to their intended impact. We believe this high quality communication of intended impact led to higher quality research, and is on the path towards what the broader community should be striving for across the board.

3 Worksheet Proposal

With the previous case studies in mind, we suggest future researchers consider a series of concrete questions about their potential research. Answering these questions will not only prompt reflection for the researchers, it will provide a critical component of transparency for future readers about the researcher's thought processes in the design phase of their research. We are inspired partially by the concept of preregistration. Whereas preregistration helps researchers distinguish between prediction and postdiction, our recommendations help prediction researchers reflect on how to maximize the intentional and minimize the unintentional downstream consequences of their research [22].

The idealized scenario for prediction research involves a researcher identifying a problem domain in which interventions would be improved with access to more accurate predictions. The researcher selects a dataset to base their predictions on, either by using a preexisting dataset or by collecting the necessary data themselves. The researcher determines how to evaluate their predictions, and selects one or more candidate mechanisms for the actual prediction. They implement their prediction mechanism over their training dataset, and evaluate its performance on the holdout dataset. They then use these predictions to improve interventions in the problem domain, or use their insights to repeat the cycle of improving predictions for that problem domain. In this idealized scenario, our recommendations focus on the initial decision by the researcher to improve interventions in a problem domain, and the ultimate deployment of predictions for interventions in that problem domain.

We recognize that many challenges intersect to produce weak research, and we do not envision our recommendations as a panacea for all these challenges. Our intention is far more humble: to provide a rigorous, consistent, and transparent process for researchers to reflect on the downstream

impact of their predictions. To that end, we suggest researchers document their responses to the following questions, and perhaps include this documentation as an appendix to their publication:

1. What impact do you intend your research to have?
2. What are some of the ways the intended impact can be achieved without predictions? Why are predictions the best suited tool for this intended impact?
3. Can you anticipate any secondary consequences of your work? How could your (accurate or inaccurate) predictions be misused?
4. What is the risk of harm towards marginalized communities from your research, according to those critical of this domain of work? To what extent does this study work to address these concerns?
5. What are the ethical issues related to this domain of research, according to those critical of this domain of work? How will these ethical concerns impact the research design?
6. To what extent have you engaged with stakeholders in this problem domain? Which stakeholders do you think are excited? Which stakeholders do you think are skeptical? For each stakeholder you've engaged with, please describe their general sentiment towards this work.
7. To what extent does the proposed work deviate from research norms in this problem domain? For each deviation, please explain what prompted this deviation?

These questions will be useless if researchers approach them as a rote chore to complete before the actual work is done. However, genuine self-reflection is a difficult process, and one that traditional computer science education does not prioritize, so guidance must be thorough. Therefore, we have tried to strike a balance between completeness and simplicity. In other words we want to ask the hard questions without coming across as intimidating, adversarial, or excessive. Because striking any balance is a deliberative process, we do not view these questions as set in stone, and we expect them to evolve over time.

4 Question Justification

We will now use the case studies from before to demonstrate the value of each question we proposed.

4.1 "What impact do you intend your research to have?"

Two different genetic researchers may have different goals, which is of course fine. One may be more interested in the implications their research will have on scientific understanding of the human body. The other may be more interested in the medical applications. Our intent in asking this question is not to filter out projects that do not meet our personal definition of "high impact". But clarifying the intended impact will benefit both the researcher themselves in tailoring their research to that goal and benefit the reader in understanding the context under which the research

was pursued. As described above, the Fragile Families Challenge was particularly successful in this regard. Committing to these intentions will also prevent the repurposing of research to fit trendy topics.

4.2 "What are some of the ways the intended impact can be achieved without predictions? Why are predictions the best suited tool for this intended impact?"

A criminal risk assessment tool for deciding which defendants should be placed in pretrial incarceration may use this space to talk about the harms of the status quo that doesn't use risk assessment but instead uses cash bail. If the criminal risk assessment researchers would like to have the intended impact of reducing the number of individuals held before their trial, this space can be used to discuss public policy approaches to the problem, such as advocating against pretrial incarceration based on the presumption of innocence before a guilty verdict is reached. This question gives researchers space to look outside their traditional domain and learn about the state of the art in other fields, which can provide multi-disciplinary depth to their work.

We also hope some researchers will find insight at this point that may lead them away from prediction altogether. This is best demonstrated by one of our own research projects studying pathways to far-right radicalization online.

The internet has provided numerous platforms for users to be exposed to, consume, and share dangerous extremist political content. In our research, we are focusing on white supremacist beliefs that can range from "alt-lite" beliefs (such as general anti-immigration sentiment) to alt-right beliefs (such as advocacy for violent ethnic cleansing). Whereas, once upon a time, such views would be tentatively shared among small, scattered communities, the internet has provided the opportunity for such extremist beliefs to metastasize in the public consciousness. This phenomenon has galvanized white supremacist violence, and led the Department of Homeland Security to characterize white supremacist extremists as "the most persistant and lethal threat" to American safety in their 2020 Homeland Threat Assessment [23].

With this in mind, we would like to quantitatively track the pathways to radicalization and the exposure to, consumption of, and sharing of far-right content through a variety of online platforms. These online platforms can range anywhere from the popular video sharing platform YouTube, to the controversial and anonymous image board 4chan, to the openly neo-Nazi disinformation outlet StormFront. We would like to learn what role each platform plays in the extremist content ecosystem. In particular: what are the comparative roles of each platform in the exposure to, consumption of, and sharing of extremist content; how is the extremist content on a platform integrated into the mainstream content on that platform; and what causes users to be exposed to the extremist content they view on a platform? Armed with this analysis, we hope policymakers and platform stakeholders will be better prepared to design interventions that curb the spread of extremist content online.

This research could be taken in a slightly different direction, however. In order to curb white supremacist violence, you could imagine designing an online tool to track the white supremacists themselves, and predict which internet users are likely to pose violent threats to public safety. We believe that intervening at the platform level, rather than intervening at the individual level more directly addresses the root cause of the problem (the spread of extremist content itself, rather than

individual cases of extremist content consumption). If asked to justify going down the predictive route, we would be forced to conclude that there are better tools for achieving our desired impact.

4.3 "Can you anticipate any secondary consequences of your work? How could your (accurate or inaccurate) predictions be misused?"

Although a researcher may have well-defined and positive intentions for the impact of their research, they may not be able to guarantee all readers share their goals. This question can prompt them to explore such misuse of knowledge. For example, Fragile Families researchers may wonder if their predictions would be used by evil insurance companies to price discriminate against vulnerable families. When considering that the criminal legal system has been weaponized against black communities since chattel slavery [1] [24] [6], criminal risk assessment researchers may worry that racist policymakers are not as invested in the technical "accuracy" of the tool as the research community may be. It is important to note, however, that we do not intend for researchers to restrict the potential misuse of their technology strictly to individuals, and hope they recognize the flawed societal structures within which their work may be used.

4.4 "What is the risk of harm towards marginalized communities from your research, according to those critical of this domain of work? To what extent does this study work to address these concerns?"

It is understandably difficult to imagine oneself as part of the problem. Instead, perhaps asking researchers to preemptively imagine what criticism might be leveled against their research will help the researchers recognize their own harmful downstream effects.

Viewing through a reproductive justice lens [28], a genetic screening researcher could recognize the genetic screens they develop could be disproportionately applied to black and brown mothers for negative eugenics. Viewing through a prison abolitionist lens [8], a risk assessment researcher could recognize that risk prediction scores scientifically legitimize the deep-seated racism of the criminal legal system. Asking researchers to consider such criticism before they are deeply invested in the research may help them recognize their complicit role in oppression.

Of course, though, we acknowledge that being asked simple questions like these is unlikely to fundamentally shift a researcher's worldview. But in both cases, prompting the researcher to seriously weigh existing criticisms may help promote interaction between two communities that traditionally speak past each other. Furthermore, if there are tangible steps the researcher can take to mitigate these dangers, it is better for them to catch those opportunities sooner rather than later.

4.5 "What are the ethical issues related to this domain of research, according to those critical of this domain of work? How will these ethical concerns impact the research design?"

The most unnerving aspect of the neo-eugenics work by Armand Marie Leroi was his explicit refusal to engage with the acknowledged ethical dimension of his work: "Although I do not doubt [the ethical or legal implications'] importance, I simply have no particular knowledge of them." [18]. Answering this question would take the researcher through the first step of having a "particular knowledge" of the ethical issues at stake. Again, we are not requiring the researcher to come

to one conclusion over another conclusion, but we would prefer the researcher to take the ethical implications of their work seriously.

4.6 "To what extent have you engaged with stakeholders in this problem domain? Which stakeholders do you think are excited? Which stakeholders do you think are skeptical? For each stakeholder you've engaged with, please describe their general sentiment towards this work."

Susan Athey has successfully outlined how "methods optimized solely for prediction also do not account for other factors that may be important in data-driven policy analysis or resource allocation" [3]. We believe stakeholders would be able to inform researchers about these factors. We recognize that engaging with stakeholders is a difficult and time consuming process, though. We hope that this question will emphasize that engaging with stakeholders may seem unnecessary or infeasible, but it will ultimately help the research achieve its intended impacts.

This question also comes from our genuine curiosity about what policymakers thought about the idea of predicting life outcomes when the Fragile Families Challenge started. The paper discusses that policymakers value quantitative insights in general, but did not address what relevant stakeholders thought in the Fragile Families domain in particular. Even if the researchers collected this information, such insights aren't relevant to every reader, which means it doesn't belong in the main paper itself. But having this worksheet as an appendix, for example, would provide transparency for curious readers about such details.

It is important to note that we intend the researchers to interpret "stakeholder" broadly. Although lawyers, judges, police departments, and the public are important stakeholders in criminal risk assessments, the impact of such assessments on the prisoners themselves should not be ignored. Once again, it can seem unnecessary and unlikely for quantitative researchers to interact directly with prisoners, for example, but that would represent a fundamental shortcoming in the research.

We recognize that, due to the logistical difficulties of engaging directly with a variety of stakeholders, it may be sufficient for researchers to consult their writings and publications instead of consulting them directly. In this case, we intend the researchers to provide their best judgement about how they expect those stakeholders to react to their research, and list what writings and publications they are basing this viewpoint on.

4.7 "To what extent does the proposed work deviate from research norms in this problem domain? For each deviation, please explain what prompted this deviation."

While studies such as Dawes et al. [9] find that actuarial methods lead to superior prediction when compared to the judgement of domain experts, this does not provide researchers with a license to abandon the norms of the field they are attempting to contribute to or disrupt. When applying predictive models within a domain, researchers should deliberately consider all of the best practices of the domain. Additionally, researchers should make a best effort to follow these norms and, if a researcher decides to deviate from domain standards, they should elucidate their reasoning for such a decision. In the development of predictive models in the healthcare diagnoses context, for example, studies should be prospective and be conducted in a clinical setting when possible

in order to prevent exaggerated claims that could present a risk to patient safety. Furthermore, standard reporting practices should be followed in order to support further scientific progress and allow for external validation of results.

5 Related Work

In *Critique and Praxis*, Bernard Harcourt describes a failure of critical philosophers to substantively engage with critical practice. He contrasts how "Many critical philosophers today—even some of the leading critical theorists of our time—now openly resist the call to praxis" with the "critical voices who have stayed true to the ambition of critical praxis" [16]. In the context of our paper, we are criticizing the prediction research community's focus on methods and techniques of prediction (the theory of prediction), and recommending that the field should refocus on the actual details that will affect real world interventions (the praxis of prediction). That is not to say we are suggesting all predictions get implemented as interventions. In fact, quite the opposite. We are suggesting that if researchers "articulate a practice or program", then the community can finally engage in a "critical debate over our own critical practices" [16]. In other words, the epistemological focus on the technical details of prediction preempt the question of how we expect the predictions to actually change the world.

In *Data Science as Political Action*, Ben Green describes how the data science community (which includes prediction researchers) has responded to criticisms about "the social harms associated with data-driven algorithms" by adopting "ethics training and principles". He posits that such efforts are "ill-equipped to address broad matters of social justice", and instead explains "why data scientists must recognize themselves as political actors" and "how the field can evolve toward a deliberative and rigorous grounding in a politics of social justice" [12]. He first addresses common defenses of data scientists about the political position of their work, then frames four stages of incorporating politics into data science: "becoming interested in directly addressing social issues, recognizing the politics underlying these issues, redirecting existing methods toward new applications that challenge oppression, and developing practices and methods for working with communities as productive partners in movements for social justice." [12]. We view the questions we pose to researchers as a jumping off point for them to explore these stages. For example, question one addresses stage one, questions four and five address stage two, and question six addresses stage four.

We are far from the first ones to raise concerns about predictions and their role in the ultimate interventions [7] [5] [4].

For example, Cederman and Weidmann describe the state of the art in predicting armed conflicts, outline some difficulties with the process, and recommend adjusting expectations of what predictions can provide as far as interventions go. They say "Scholars producing forecasts typically assume that policymakers want predictive risk assessments more than anything else because this would allow them to reduce potential conflict through preventive resource allocation and intervention. However, these hopes presuppose that the effects of policy intervention are well known. [...] Given the difficulties of obtaining reliable information on key social indicators, especially in developing countries, basic description and explanatory modeling may, in many instances, be more urgently needed than forecasting." [7]. We hope that our sixth questions will prompt such scholars

to engage with the policymakers they aim to serve, and we hope that our second question may prompt them to consider basic description and explanatory modeling rather than forecasting.

Similarly, Chelsea Barabas et al. also suggest reframing the field to prioritize *Interventions over Predictions*. In the criminal risk legal system context, they "outline key differences between regression, machine learning and causal inference in order to make the case for moving away from using regression and machine learning for intervention-oriented assessment [5]". They "argue that when risk assessments are used primarily as a predictive technology, they fuel harmful trends towards mass incarceration and growing inequality in the justice system.", which follows our line of reasoning that intended and actual downstream effects of research may not line up.

Taking a broader view of big data applied to policy problems in general (rather than focusing on any specific context), Susan Athey goes on to point out how "there are a number of gaps between making a prediction and making a decision, and underlying assumptions need to be understood in order to optimize data-driven decision-making." [4]. This paper provides a good starting point for researchers to recognize the needs of policymakers, and we believe that engagement with stakeholders directly in a researcher's problem domain will bear further fruit.

As discussed at the beginning of section 3, we view our work in line with that of preregistering hypotheses [22]. Whereas preregistration helps researchers distinguish between prediction and postdiction, our recommendations help prediction researchers reflect on how to maximize the intentional and minimize the unintentional downstream consequences of their research. Although we are recommending a worksheet of reflection questions for researchers to contemplate, we do not intend the regulation of research based on specific responses, as happened with Internal Review Boards and the National Research Act of 1974 after the publication of the Belmont Report [10].

6 Conclusion

We will conclude where we began: with Philip Rogaway's call for cryptographers (and computer scientists more generally) to reevaluate the moral character of their work. He warns "you can ignore this landscape of power, and all political and moral dimensions of our field. But that won't make them go away. It will just tend to make your work less relevant or socially useful." [25]. In our paper, we used four case studies to demonstrate how prediction researchers can implement our recommendations to make their work more relevant and socially useful, while avoiding unintended secondary consequences. Although our recommendations are by no means a panacea to the problems we identified, we hope our worksheet provides concrete recommendations for reflective practice.

References

[1] Michelle Alexander. 2020. *The new Jim Crow: Mass incarceration in the age of colorblindness*. The New Press.

[2] Julia Angwin and Jeff Larson. 2016. Bias in criminal risk scores is mathematically inevitable, researchers say. *Propublica, available at: https://goo. gl/S3Gwcn (accessed 5 March 2018)* (2016).

[3] Susan Athey. 2017a. Beyond prediction: Using big data for policy problems. *Science* 355, 6324 (2017), 483–485. DOI:http://dx.doi.org/10.1126/science.aal4321

[4] Susan Athey. 2017b. Beyond prediction: Using big data for policy problems. *Science* 355, 6324 (2017), 483–485. DOI:http://dx.doi.org/10.1126/science.aal4321

[5] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. 2018. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on Fairness, Accountability and Transparency*. PMLR, 62–76.

[6] Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity Press.

[7] Lars-Erik Cederman and Nils B. Weidmann. 2017. Predicting armed conflict: Time to adjust our expectations? *Science* 355, 6324 (2017), 474–476. DOI:http://dx.doi.org/10.1126/science.aal4483

[8] Angela Y. Davis. 2011. *Are prisons obsolete?* Seven Stories Press.

[9] RM Dawes, D Faust, and PE Meehl. 1989. Clinical versus actuarial judgment. *Science* 243, 4899 (1989), 1668–1674. DOI:http://dx.doi.org/10.1126/science.2648573

[10] Education Department of Health and others. 2014. The Belmont Report. Ethical principles and guidelines for the protection of human subjects of research. *The Journal of the American College of Dentists* 81, 3 (2014), 4.

[11] Léon Digard and Elizabeth Swavola. 2019. Justice denied: The harmful and lasting effects of pretrial detention. *Vera Evidence Brief. New York: Vera Institute of Justice* (2019).

[12] Ben Green. 2020a. Data science as political action: grounding data science in a politics of justice. *Available at SSRN 3658431* (2020).

[13] Ben Green. 2020b. The false promise of risk assessments: epistemic reform and the limits of fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 594–606.

[14] Varun Gulshan, Lily Peng, Marc Coram, Martin Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip Nelson, Jessica Mega, and Dale Webster. 2016. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* 316 (11 2016). DOI:http://dx.doi.org/10.1001/jama.2016.17216

[15] Benjamin Haibe-Kains, George Adam, Ahmed Hosny, Farnoosh Khodakarami, MAQC Board, Levi Waldron, Bo Wang, Chris Mcintosh, Anshul Kundaje, Casey Greene, Michael Hoffman, Jeffrey Leek, Wolfgang Huber, Alvis Brazma, Joelle Pineau, Robert Tibshirani, Trevor Hastie, John Ioannidis, John Quackenbush, and Hugo Aerts. 2020. The importance of transparency and reproducibility in artificial intelligence research. (2020). https://doi.org/10.1038/s41586-020-2766-y

[16] Bernard E. Harcourt. 2020. *The Primacy of Critique and Praxis*. Columbia University Press, 1–18. http://www.jstor.org/stable/10.7312/harc19572.3

[17] Will Douglas Heaven. 2020. AI is wrestling with a replication crisis. *MIT Technology Review* (2020).

[18] Armand Marie Leroi. 2006. The future of neo-eugenics: Now that many people approve the elimination of certain genetically defective fetuses, is society closer to screening all fetuses for all known mutations? *EMBO reports* 7, 12 (2006), 1184–1187.

[19] Randall Lewis, Justin M Rao, and David H Reiley. 2013. *Measuring the Effects of Advertising: The Digital Frontier*. Working Paper 19520. National Bureau of Economic Research. DOI:http://dx.doi.org/10.3386/w19520

[20] Scott McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher Kelly, Dominic King, and Shravya Shetty. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577 (01 2020), 89–94. DOI:http://dx.doi.org/10.1038/s41586-019-1799-6

[21] Myura Nagendran, Yang Chen, Christopher A Lovejoy, Anthony C Gordon, Matthieu Komorowski, Hugh Harvey, Eric J Topol, John P A Ioannidis, Gary S Collins, and Mahiben Maruthappu. 2020. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 368 (2020). DOI:http://dx.doi.org/10.1136/bmj.m689

[22] Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences* 115, 11 (2018), 2600–2606. DOI:http://dx.doi.org/10.1073/pnas.1708274114

[23] Department of Homeland Security. 2020. *Homeland Threat Assessment*. Technical Report. https://www.dhs.gov/sites/default/files/publications/2020_10_06_homeland-threat-assessment.pdf

[24] Dorothy E Roberts. 2019. Abolition Constitutionalism. *Harv. L. Rev.* 133 (2019), 1.

[25] Phillip Rogaway. 2016. The Moral Character of Cryptographic Work. USENIX Association, Austin, TX.

[26] Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, Jennie E. Brand, Nicole Bohme Carnegie, Ryan James Compton, Debanjan Datta, Thomas Davidson, Anna Filippova, Connor Gilroy, Brian J. Goode, Eaman Jahani, Ridhi Kashyap, Antje Kirchner, Stephen McKay, Allison C. Morgan, Alex Pentland, Kivan Polimis, Louis Raes, Daniel E. Rigobon, Claudia V. Roberts, Diana M. Stanescu, Yoshihiko Suhara, Adaner Usmani, Erik H. Wang, Muna Adem, Abdulla Alhajri, Bedoor AlShebli, Redwane Amin, Ryan B. Amos, Lisa P. Argyle, Livia Baer-Bositis, Moritz Büchi, Bo-Ryehn Chung, William Eggert, Gregory Faletto, Zhilin

Fan, Jeremy Freese, Tejomay Gadgil, Josh Gagné, Yue Gao, Andrew Halpern-Manners, Sonia P. Hashim, Sonia Hausen, Guanhua He, Kimberly Higuera, Bernie Hogan, Ilana M. Horwitz, Lisa M. Hummel, Naman Jain, Kun Jin, David Jurgens, Patrick Kaminski, Areg Karapetyan, E. H. Kim, Ben Leizman, Naijia Liu, Malte Möser, Andrew E. Mack, Mayank Mahajan, Noah Mandell, Helge Marahrens, Diana Mercado-Garcia, Viola Mocz, Katariina Mueller-Gastell, Ahmed Musse, Qiankun Niu, William Nowak, Hamidreza Omidvar, Andrew Or, Karen Ouyang, Katy M. Pinto, Ethan Porter, Kristin E. Porter, Crystal Qian, Tamkinat Rauf, Anahit Sargsyan, Thomas Schaffner, Landon Schnabel, Bryan Schonfeld, Ben Sender, Jonathan D. Tang, Emma Tsurkov, Austin van Loon, Onur Varol, Xiafei Wang, Zhi Wang, Julia Wang, Flora Wang, Samantha Weissman, Kirstie Whitaker, Maria K. Wolters, Wei Lee Woon, James Wu, Catherine Wu, Kengran Yang, Jingwen Yin, Bingyu Zhao, Chenyun Zhu, Jeanne Brooks-Gunn, Barbara E. Engelhardt, Moritz Hardt, Dean Knox, Karen Levy, Arvind Narayanan, Brandon M. Stewart, Duncan J. Watts, and Sara McLanahan. 2020. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences* 117, 15 (2020), 8398–8403. DOI: http://dx.doi.org/10.1073/pnas.1915006117

[27] Mike Voets, Kajsa Møllersen, and Lars Ailo Bongo. 2019. Reproduction study using public data of: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *PLoS ONE* 14, 6 (June 2019), e0217541. DOI: http://dx.doi.org/10.1371/journal.pone.0217541

[28] B. D. Welch. 2019. *An Ethical Analysis of Reproductive Justice in the Context of the Eugenics Movement in the United States*. Ph.D. Dissertation. https://search.proquest.com/dissertations-theses/ethical-analysis-reproductive-justice-context/docview/2234775419/se-2?accountid=13314