



The Complex Systems View of AI Ethics

Tina Eliassi-Rad

RADLAB

Trustworthy Network Science

Adversarial Robustness

Sensitivity & Stability

Explainability

Fairness & Equity

Reliability

Topological interpretations

Geometric interpretations

Just Machine Learning

Data issues

Model issues

Task issues

Community issues

Democratic backsliding

Impacts on high stakes situations

Representation issues

RADLAB

Trustworthy Network Science

Adversarial Robustness

Sensitivity & Stability

Explainability

Fairness & Equity

Reliability

Topological interpretations

Geometric interpretations

Just Machine Learning

Data issues

Model issues

Task issues

Community issues

Democratic backsliding

Impacts on high stakes situations

Representation issues



Machine learning (ML) systems are not islands.

They are part of broader complex systems.

To understand and mitigate the risks and harms of using ML, we must remove our optimization blinders & study the broader complex systems in which ML systems operate.



Age of Prediction



A prediction task

"At a given risk score, Black patients are considerably sicker than White patients, ... The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients."

what they were measuring ≠ what they thought they were measuring

RESEARCH

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer^{1,2}*, Brian Powers³, Christine Vogeli⁴, Sendhil Mullainathan⁵*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

researcher-created algorithms (10-13), With-

out an algorithm's training data, objective func-

tion, and prediction methodology, we can only

guess as to the actual mechanisms for the

important algorithmic disparities that arise.

In this study, we exploit a rich dataset that

provides insight into a live, scaled algorithm

deployed nationwide today. It is one of the

largest and most typical examples of a class

of commercial risk-prediction tools that, by

industry estimates, are applied to roughly

200 million people in the United States each

year. Large health systems and payers rely on

this algorithm to target patients for "high-risk

care management" programs. These programs

seek to improve the care of patients with

complex health needs by providing additional

resources, including greater attention from

trained providers, to help ensure that care is

well coordinated. Most health systems use

these programs as the cornerstone of pop-

ulation health management efforts, and they

are widely considered effective at improving

outcomes and satisfaction while reducing costs (14-17). Because the programs are themselves

expensive-with costs going toward teams of

dedicated nurses, extra primary care appoint-

ment slots, and other scarce resources-health

systems rely extensively on algorithms to iden-

tify patients who will benefit the most (18, 19).

challenging causal inference problem that

requires estimation of individual treatment ef-

fects. To solve this problem, health systems

make a key assumption: Those with the great-

est care needs will benefit the most from the

program. Under this assumption, the targeting

problem becomes a pure prediction policy prob-

lem (20). Developers then build algorithms

Identifying patients who will derive the greatest benefit from these programs is a

here is growing concern that algorithms may reproduce racial and gender disparities via the people building them or through the data used to train them (1-3). Empirical work is increasingly lending support to these concerns. For example, job search ads for highly paid positions are less likely to be presented to women (4), searches for distinctively Black-sounding names are more likely to trigger ads for arrest records (5), and image searches for professions such as CEO produce fewer images of women (6). Facial recognition systems increasingly used in law enforcement perform worse on recognizing faces of women and Black individuals (7, 8), and natural language processing algorithms encode language in gendered ways (9). Empirical investigations of algorithmic blas. though, have been hindered by a key constraint: Algorithms deployed on large scales are typically proprietary, making it difficult for independent researchers to dissect them. Instead, researchers must work "from the outside," often with great ingenuity, and resort to clever workarounds such as audit studies. Such efforts can document disparities, but understanding how and why they arise-much less figuring out what to do about them-is difficult without greater access to the algorithms themselves. Our understanding of a mechanism therefore typically relies on theory or exercises with

School of Public Heatti, University of California, Berkeley, Berkeley, CA, USA. "Department of Emigency Medicine Brigham and Women's Nogalik, Booton, MA, USA, Siala, Department of Medicine, Brigham and Women's Hospital, Bestnin, MA, USA. "Mongan Institute Health Policy Conter: Massachusetts General Hospital, Boston, MA, USA. "Booth School of Business, University of Chicago, Chicago IL, USA. "These authors confributed sequily to its work. "Chicagobooth.edu that rely on past data to build a predictor of future health care needs.

Our dataset describes one such typical algorithm. It contains both the algorithm's predictions as well as the data needed to understand its inner workings; that is, the underlying ingredients used to form the algorithm (data, objective function, etc.) and links to a rich set of outcome data. Because we have the inputs, outputs, and eventual outcomes, our data allow us a rare opportunity to quantify racial disparities in algorithms and isolate the mechanisms by which they arise. It should be emphasized that this algorithm is not unique. Rather, it is emblematic of a generalized approach to risk prediction in the health sector, widely adopted by a range of for- and non-profit medical centers and governmental agencies (21).

Our analysis has implications beyond what we learn about this particular algorithm. First, the specific problem solved by this algorithm has analogies in many other sectors: The predicted risk of some future outcome (in our case, health care needs) is widely used to target policy interventions under the assumption that the treatment effect is monotonic in that risk, and the methods used to build the algorithm are standard. Mechanisms of bias uncovered in this study likely operate elsewhere. Second, even beyond our particular finding, we hope that this exercise illustrates the importance, and the large opportunity, of studying algorithmic bias in health care, not just as a model system but also in its own right. By any standard-e.g., number of lives affected, life-and-death consequences of the decisionhealth is one of the most important and widespread social sectors in which algorithms are already used at scale today, unbeknownst to many.

Data and analytic strategy

Working with a large academic hospital, we identified all primary care patients enrolled in risk-based contracts from 2013 to 2015. Our primary interest was in studying differences between White and Black patients. We formed race categories by using hospital records, which are based on patient self-reporting. Any patient who identified as Black was considered to be Black for the purpose of this analysis. Of the remaining patients, those who self-identified as races other than White (e.g., Hispanic) were so considered (data on these patients are presented in table S1 and fig. S1 in the supplementary materials). We considered all remaining patients to be White. This approach allowed us to study one particular racial difference of social and historical interest between patients who self-identified as Black and patients who self-identified as White without another race or ethnicity; it has the disadvantage of not allowing for the study of intersectional racial

The prediction task

"At a given risk score, Black patients are considerably sicker than White patients, ... The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients."

RESEARCH

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer^{1,2}*, Brian Powers³, Christine Vogeli⁴, Sendhil Mullainathan⁵*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

researcher-created algorithms (10-13), With-

out an algorithm's training data, objective func-

tion, and prediction methodology, we can only

guess as to the actual mechanisms for the

important algorithmic disparities that arise.

In this study, we exploit a rich dataset that

provides insight into a live, scaled algorithm

deployed nationwide today. It is one of the

largest and most typical examples of a class

of commercial risk-prediction tools that, by

industry estimates, are applied to roughly

200 million people in the United States each

year. Large health systems and payers rely on

this algorithm to target patients for "high-risk

care management" programs. These programs

seek to improve the care of patients with

complex health needs by providing additional

resources, including greater attention from

trained providers, to help ensure that care is

well coordinated. Most health systems use

these programs as the cornerstone of pop-

ulation health management efforts, and they

are widely considered effective at improving

outcomes and satisfaction while reducing costs

(14-17). Because the programs are themselves

expensive-with costs going toward teams of

dedicated nurses, extra primary care appoint-

ment slots, and other scarce resources-health

systems rely extensively on algorithms to iden-

tify patients who will benefit the most (18, 19).

challenging causal inference problem that

requires estimation of individual treatment ef-

fects. To solve this problem, health systems

make a key assumption: Those with the great-

est care needs will benefit the most from the

program. Under this assumption, the targeting

problem becomes a pure prediction policy prob-

lem (20). Developers then build algorithms

Identifying patients who will derive the greatest benefit from these programs is a

here is growing concern that algorithms may reproduce racial and gender disparities via the people building them or through the data used to train them (1-3). Empirical work is increasingly lending support to these concerns. For example, job search ads for highly paid positions are less likely to be presented to women (4), searches for distinctively Black-sounding names are more likely to trigger ads for arrest records (5), and image searches for professions such as CEO produce fewer images of women (6). Facial recognition systems increasingly used in law enforcement perform worse on recognizing faces of women and Black individuals (7, 8), and natural language processing algorithms encode language in gendered ways (9). Empirical investigations of algorithmic blas. though, have been hindered by a key constraint: Algorithms deployed on large scales are typically proprietary, making it difficult for independent researchers to dissect them. Instead, researchers must work "from the outside," often with great ingenuity, and resort to clever workarounds such as audit studies. Such efforts can document disparities, but understanding how and why they arise-much less figuring out what to do about them-is difficult without greater access to the algorithms themselves. Our understanding of a mechanism therefore typically relies on theory or exercises with

¹School of Fubile Health, University of Galidoma, Barkelay, Barkelay, CA, USA. ³Department of Emergency Medicine, Brightm and Woman's Hospital, Boston, MA, USA. ³Department of Medicine, Brightam and Women's Hospital, Boston, MA, USA. ⁴Mongan Institute Health Policy Center, Massachusetts General Hospital, Boston, MA, USA. ⁸Booth School of Business, University of Chicogo, Chicogo, IL, USA. ⁴These audross controllated equility to this work, ⁴Corresponding author, Email: sendhilumullainsthantis chicagobooth. Adu. that rely on past data to build a predictor of future health care needs.

Our dataset describes one such typical algorithm. It contains both the algorithm's predictions as well as the data needed to understand its inner workings; that is, the underlying ingredients used to form the algorithm (data, objective function, etc.) and links to a rich set of outcome data. Because we have the inputs, outputs, and eventual outcomes, our data allow us a rare opportunity to quantify racial disparities in algorithms and isolate the mechanisms by which they arise. It should be emphasized that this algorithm is not unique. Rather, it is emblematic of a generalized approach to risk prediction in the health sector, widely adopted by a range of for- and non-profit medical centers and governmental agencies (21).

Our analysis has implications beyond what we learn about this particular algorithm. First, the specific problem solved by this algorithm has analogies in many other sectors: The predicted risk of some future outcome (in our case, health care needs) is widely used to target policy interventions under the assumption that the treatment effect is monotonic in that risk, and the methods used to build the algorithm are standard. Mechanisms of bias uncovered in this study likely operate elsewhere. Second, even beyond our particular finding, we hope that this exercise illustrates the importance, and the large opportunity, of studying algorithmic bias in health care, not just as a model system but also in its own right. By any standard-e.g., number of lives affected, life-and-death consequences of the decisionhealth is one of the most important and widespread social sectors in which algorithms are already used at scale today, unbeknownst to many.

Data and analytic strategy

Working with a large academic hospital, we identified all primary care patients enrolled in risk-based contracts from 2013 to 2015. Our primary interest was in studying differences between White and Black patients. We formed race categories by using hospital records, which are based on patient self-reporting. Any patient who identified as Black was considered to be Black for the purpose of this analysis. Of the remaining patients, those who self-identified as races other than White (e.g., Hispanic) were so considered (data on these patients are presented in table S1 and fig. S1 in the supplementary materials). We considered all remaining patients to be White. This approach allowed us to study one particular racial difference of social and historical interest between patients who self-identified as Black and patients who self-identified as White without another race or ethnicity; it has the disadvantage of not allowing for the study of intersectional racial

Risk assessment: A popular prediction task

- Why is it popular?
 - Because machine learning people know a lot about it i.e., it is a "low-hanging fruit".
 - 2. Because human decision-makers like the output of these tasks; they are easily understandable.



Issues with risk assessment





Parity (a.k.a. fairness) measures come from the confusion (a.k.a. error) matrix

		True condition			
	Total population	Condition positive	Condition negative		
Predicted condition	Predicted condition positive	True positive	False positive, Type I error		
	Predicted condition negative	False negative, Type II error	True negative		

Image from https://en.wikipedia.org/wiki/Confusion matrix

Parity (a.k.a. fairness) measures come from the confusion (a.k.a. error) matrix

		True co	ndition			
	Total population	Condition positive	Condition negative	$\frac{\text{Prevalence}}{\sum \text{ Condition positive}} = \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Acc Σ True posi Σ Tr	tive + $Σ$ True negative total population
condition	Predicted condition positive	True positive	False positive, Type I error	Positive predictive value (PPV), Precision = Σ True positive $\overline{\Sigma}$ Predicted condition positive	False dis Σ Ι Σ Predicto	covery rate (FDR) = False positive ed condition positive
Predicted	Predicted condition negative	False negative, Type II error	True negative	False omission rate (FOR) = Σ False negative Σ Predicted condition negative	Negative predictive value (NPV) = Σ True negativeΣ Predicted condition negative	
		True positive rate (TPR),Recall, Sensitivity,probability of detection,Power = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR),Fall-out,probability of false alarm Σ False positive Σ Condition negative	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOB)	F ₁ score =
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR-) = FNR TNR	$=\frac{LR+}{LR-}$	2 · Precision + Recall

Image from https://en.wikipedia.org/wiki/Confusion matrix

Condition	Chouldechova (2016)
Mutual exclusivity between groups a and b	\checkmark
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	\checkmark

Condition	Chouldechova (2016)
Mutual exclusivity between groups a and b	\checkmark
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	\checkmark
Imperfect classifier: $Pr_a(C = 1 Y = 0) \neq 0$ and $Pr_b(C = 1 Y = 0) \neq 0$ and $Pr_a(C = 1 Y = 1) \neq 1$ and $Pr_b(C = 1 Y = 1) \neq 1$	
Non-zero precision: $Pr_a(Y = 1 C = 1) \neq 0$ <u>or</u> $Pr_b(Y = 1 C = 1) \neq 0$	
Statistical parity: $Pr_a(C = 1) = Pr_b(C = 1)$	\checkmark

Condition	Chouldechova (2016)
Mutual exclusivity between groups a and b	\checkmark
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	\checkmark
Imperfect classifier: $Pr_a(C = 1 Y = 0) \neq 0$ and $Pr_b(C = 1 Y = 0) \neq 0$ and $Pr_a(C = 1 Y = 1) \neq 1$ and $Pr_b(C = 1 Y = 1) \neq 1$	
Non-zero precision: $Pr_a(Y = 1 C = 1) \neq 0$ <u>or</u> $Pr_b(Y = 1 C = 1) \neq 0$	
Statistical parity: $Pr_a(C = 1) = Pr_b(C = 1)$	\checkmark
Predictive parity: $Pr_a(Y = 1 C = 1) = Pr_b(Y = 1 C = 1)$	×
True positive parity: $Pr_a(C = 1 Y = 1) = Pr_b(C = 1 Y = 1)$	×
False positive parity: $Pr_a(C = 1 Y = 0) = Pr_b(C = 1 Y = 0)$	×

Condition	Chouldechova (2016)	Kleinberg et al. (2016)
Mutual exclusivity between groups a and b	\checkmark	\checkmark
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	\checkmark	\checkmark
Imperfect classifier: $Pr_a(C = 1 Y = 0) \neq 0$ and $Pr_b(C = 1 Y = 0) \neq 0$ and $Pr_a(C = 1 Y = 1) \neq 1$ and $Pr_b(C = 1 Y = 1) \neq 1$		
Non-zero precision: $Pr_a(Y = 1 C = 1) \neq 0$ <u>or</u> $Pr_b(Y = 1 C = 1) \neq 0$		
Statistical parity: $Pr_a(C = 1) = Pr_b(C = 1)$	\checkmark	
Predictive parity: $Pr_a(Y = 1 C = 1) = Pr_b(Y = 1 C = 1)$	×	
True positive parity: $Pr_a(C = 1 Y = 1) = Pr_b(C = 1 Y = 1)$	×	
False positive parity: $Pr_a(C = 1 Y = 0) = Pr_b(C = 1 Y = 0)$	×	

Condition	Chouldechova (2016)	Kleinberg et al. (2016)
Mutual exclusivity between groups a and b	\checkmark	\checkmark
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	\checkmark	\checkmark
Imperfect classifier: $Pr_a(C = 1 Y = 0) \neq 0$ and $Pr_b(C = 1 Y = 0) \neq 0$ and $Pr_a(C = 1 Y = 1) \neq 1$ and $Pr_b(C = 1 Y = 1) \neq 1$		\checkmark
Non-zero precision: $Pr_a(Y = 1 C = 1) \neq 0$ <u>or</u> $Pr_b(Y = 1 C = 1) \neq 0$		\checkmark
Statistical parity: $Pr_a(C = 1) = Pr_b(C = 1)$	\checkmark	
Predictive parity: $Pr_a(Y = 1 C = 1) = Pr_b(Y = 1 C = 1)$	×	
True positive parity: $Pr_a(C = 1 Y = 1) = Pr_b(C = 1 Y = 1)$	×	
False positive parity: $Pr_a(C = 1 Y = 0) = Pr_b(C = 1 Y = 0)$	×	

Condition	Chouldechova (2016)	Kleinberg et al. (2016)
Mutual exclusivity between groups a and b	\checkmark	\checkmark
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	\checkmark	\checkmark
Imperfect classifier: $Pr_a(C = 1 Y = 0) \neq 0$ and $Pr_b(C = 1 Y = 0) \neq 0$ and $Pr_a(C = 1 Y = 1) \neq 1$ and $Pr_b(C = 1 Y = 1) \neq 1$		\checkmark
Non-zero precision: $Pr_a(Y = 1 C = 1) \neq 0$ <u>or</u> $Pr_b(Y = 1 C = 1) \neq 0$		\checkmark
Statistical parity: $Pr_a(C = 1) = Pr_b(C = 1)$	\checkmark	
Predictive parity: $Pr_a(Y = 1 C = 1) = Pr_b(Y = 1 C = 1)$	×	×
True positive parity: $Pr_a(C = 1 Y = 1) = Pr_b(C = 1 Y = 1)$	×	×
False positive parity: $Pr_a(C = 1 Y = 0) = Pr_b(C = 1 Y = 0)$	×	×

Condition	Chouldechova (2016)	Kleinberg et al. (2016)	Eliassi-Rad & Fitelson (2021)
Mutual exclusivity between groups a and b	\checkmark	\checkmark	
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	\checkmark	\checkmark	\checkmark
Imperfect classifier: $Pr_a(C = 1 Y = 0) \neq 0$ and $Pr_b(C = 1 Y = 0) \neq 0$ and $Pr_a(C = 1 Y = 1) \neq 1$ and $Pr_b(C = 1 Y = 1) \neq 1$		\checkmark	
Non-zero precision: $Pr_a(Y = 1 C = 1) \neq 0$ <u>or</u> $Pr_b(Y = 1 C = 1) \neq 0$		\checkmark	
Statistical parity: $Pr_a(C = 1) = Pr_b(C = 1)$	\checkmark		
Predictive parity: $Pr_a(Y = 1 C = 1) = Pr_b(Y = 1 C = 1)$	×	×	
True positive parity: $Pr_a(C = 1 Y = 1) = Pr_b(C = 1 Y = 1)$	×	×	
False positive parity: $Pr_a(C = 1 Y = 0) = Pr_b(C = 1 Y = 0)$	×	×	

Condition	Chouldechova (2016)	Kleinberg et al. (2016)	Eliassi-Rad & Fitelson (2021)
Mutual exclusivity between groups a and b	\checkmark	\checkmark	
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	\checkmark	\checkmark	\checkmark
Imperfect classifier: $Pr_a(C = 1 Y = 0) \neq 0$ and $Pr_b(C = 1 Y = 0) \neq 0$ and $Pr_a(C = 1 Y = 1) \neq 1$ and $Pr_b(C = 1 Y = 1) \neq 1$		\checkmark	\checkmark
Non-zero precision: $Pr_a(Y = 1 C = 1) \neq 0$ <u>or</u> $Pr_b(Y = 1 C = 1) \neq 0$		\checkmark	\checkmark
Statistical parity: $Pr_a(C = 1) = Pr_b(C = 1)$	\checkmark		
Predictive parity: $Pr_a(Y = 1 C = 1) = Pr_b(Y = 1 C = 1)$	×	×	
True positive parity: $Pr_a(C = 1 Y = 1) = Pr_b(C = 1 Y = 1)$	×	×	
False positive parity: $Pr_a(C = 1 Y = 0) = Pr_b(C = 1 Y = 0)$	×	×	

Condition	Chouldechova (2016)	Kleinberg et al. (2016)	Eliassi-Rad & Fitelson (2021)
Mutual exclusivity between groups a and b	\checkmark	\checkmark	
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	\checkmark	\checkmark	\checkmark
Imperfect classifier: $Pr_a(C = 1 Y = 0) \neq 0$ and $Pr_b(C = 1 Y = 0) \neq 0$ and $Pr_a(C = 1 Y = 1) \neq 1$ and $Pr_b(C = 1 Y = 1) \neq 1$		\checkmark	\checkmark
Non-zero precision: $Pr_a(Y = 1 C = 1) \neq 0$ <u>or</u> $Pr_b(Y = 1 C = 1) \neq 0$		\checkmark	\checkmark
Statistical parity: $Pr_a(C = 1) = Pr_b(C = 1)$	\checkmark		
Predictive parity: $Pr_a(Y = 1 C = 1) = Pr_b(Y = 1 C = 1)$	×	×	×
True positive parity: $Pr_a(C = 1 Y = 1) = Pr_b(C = 1 Y = 1)$	×	×	×
False positive parity: $Pr_a(C = 1 Y = 0) = Pr_b(C = 1 Y = 0)$	×	×	×

Condition	Kleinberg et al. (2016)	Eliassi-Rad & Fitelson (2021)	Eliassi-Rad & Fitelson (2021)
Mutual exclusivity between groups a and b	\checkmark		
Unequal base rates: $Pr_a(Y = 1) \neq Pr_b(Y = 1)$	\checkmark	\checkmark	\checkmark
Imperfect classifier: $Pr_a(C = 1 Y = 0) \neq 0$ and $Pr_b(C = 1 Y = 0) \neq 0$ and $Pr_a(C = 1 Y = 1) \neq 1$ and $Pr_b(C = 1 Y = 1) \neq 1$	\checkmark	\checkmark	
Non-zero precision: $Pr_a(Y = 1 C = 1) \neq 0$ <u>or</u> $Pr_b(Y = 1 C = 1) \neq 0$	\checkmark	\checkmark	
Regular priors: Prior Pr(•) only assigns extremal probability to non- contingent propositions			\checkmark
Statistical parity: $Pr_a(C = 1) = Pr_b(C = 1)$			
Predictive parity: $Pr_a(Y = 1 C = 1) = Pr_b(Y = 1 C = 1)$	×	×	×
True positive parity: $Pr_a(C = 1 Y = 1) = Pr_b(C = 1 Y = 1)$	×	\times	×
False positive parity: $Pr_a(C = 1 Y = 0) = Pr_b(C = 1 Y = 0)$	×	×	×

There are more impossibility results for risk assessment and group fairness

- To read about more impossibilities, download http://fitelson.org/exploring_impossibility.pdf
- To find new ones, download the Mathematica notebook at <u>http://fitelson.org/exploring_impossibility.nb</u>



Branden Fitelson

Fallout from the impossibility theorems

- Get rid of one of the parities
- Put bounds on the parities
- Deborah Hellman (University of Virginia Law School)
 - Predictive parity captures "what you ought to believe"
 - True positive and false positive parities capture "what you ought to do"
 - The algorithm ought not be thinking about the right-making properties when deliberating in many cases
 - → If you are going to drop a parity, drop predictive parity

Predictive parity: $Pr_a(Y = 1 | C = 1) = Pr_b(Y = 1 | C = 1)$ True positive parity: $Pr_a(C = 1 | Y = 1) = Pr_b(C = 1 | Y = 1)$ False positive parity: $Pr_a(C = 1 | Y = 0) = Pr_b(C = 1 | Y = 0)$

What about individual fairness?

- Dwork et al. [2012]: Similar individuals should be treated similarly.
- Will Fleisher. What's fair about individual fairness? AIES 2021: 480-490.
- Four problems for individual fairness as a definition and as a method for ensuring fairness:
 - 1. Insufficiency of similar treatment
 - 2. Systematic bias and arbiters (implicit human biases)
 - 3. Prior moral judgments about task relevance and moral values
 - 4. Incommensurability
- Takeaways:
 - Individual fairness is inadequate as a definition of fairness.
 - Individual fairness should not be used as a sole means for determining fairness (or detecting bias).



Will Fleisher



Image courtesy of Jennifer Wortman Vaughan

Loss Function

Learning Algorithm

26

https://www.nejm.org/doi/full/10.1056/NEJMc2029240 (Dec 2020)

Oximeter Data

CORONAVIRUS GUIDE

Devices Used In COVID-19 Treatment Can Give Errors For Patients With Dark Skin

December 16, 2020 · 5:01 PM ET





A paramedic uses a pulse oximeter to check a patient's vital signs during an August home visit in the Bronx borough of New Vork

Angus Mordant/Bloomberg via Getty Images



The NEW ENGLAND JOURNAL of MEDICINE

CORRESPONDENCE



Racial Bias in Pulse Oximetry Measurement

TO THE EDITOR: Oxygen is among the most fre- We analyzed 10,789 pairs of measures of oxyment is unknown.

analyses to measures of arterial blood gas that included carboxyhemoglobin and methemoglobin saturations.

We tested for occult hypoxemia (i.e., an arterial oxygen saturation of <88% despite an oxygen saturation of 92 to 96% on pulse oximetry) among patients who identified their race as Black or White, Since a low level of peripheral perfusion could lower the accuracy of oxygen saturation values," we also estimated the percentage of patients with occult hypoxemia after adjusting for age, sex, and cardiovascular score on the Sequential Organ Failure Assessment (SOFA) in the University of Michigan cohort. Additional details regarding the methods that were used in the study are provided in the Supplementary Appendix, available with the full text of this letter at NEJM.org.

quently administered medical therapies, with a gen saturation by pulse oximetry and arterial level that is commonly adjusted according to the oxygen saturation in arterial blood gas obtained reading on a pulse oximeter that measures pa- from 1333 White patients and 276 Black patients tients' oxygen saturation. Questions about pulse in the University of Michigan cohort and 37,308 oximeter technology have been raised, given its pairs obtained from 7342 White patients and original development in populations that were 1050 Black patients in the multicenter cohort. In not racially diverse.12 The clinical significance of the University of Michigan cohort, among the potential racial bias in pulse oximetry measure- patients who had an oxygen saturation of 92 to 96% on pulse oximetry, an arterial oxygen satu-

Our study involved adult inpatients who were ration of less than 88% was found in 88 of 749 receiving supplemental oxygen at the University arterial blood gas measurements in Black patients of Michigan Hospital (from January through July (11.7%; 95% confidence interval [CI], 8.5 to 16.0) 2020) and patients in intensive care units at 178 and in 99 of 2778 measurements in White pahospitals (from 2014 through 2015).3 We analyzed tients (3.6%; 95% CI, 2.7 to 4.7) (Fig. 1). The paired pulse oximetry measures of oxygen satu- findings in the adjusted analyses were similar to ration and measures of arterial oxygen satura- those in the unadjusted analyses, with an artetion in arterial blood gas, with all evaluations rial blood gas oxygen saturation of less than 88% performed within 10 minutes of each other. To in 11.4% (95% Cl, 7.6 to 15.2) of the measureensure that the arterial oxygen saturation was ments in Black patients and in 3.6% (95% CI, 2.5 directly measured by co-oximetry, we limited to 4.6) of those in White patients. Results were

THIS WEEK'S LETTERS

- 2477 Racial Bias in Pulse Oximetry Measurement
- Risk Factors for SARS-CoV-2 in a Statewide 2479 **Correctional System**
- 2480 Capping or Suctioning for Tracheostomy Decannulation
- 2482 Avelumab Maintenance for Urothelial Carcinoma
- 2484 A Randomized Trial of Closed-Loop Control in Children with Type 1 Diabetes
- 2485 Spectrum of Fibrotic Lung Diseases
- 2486 Racial Disproportionality in Covid Clinical Trials

N ENGL | MED 383;25 NEIM.ORG DECEMBER 17, 2020

The New England Journal of Medicine Downloaded from neum org on June 4, 2021. For personal use only, No other uses without permission. Copyright @ 2020 Massachusetts Medical Society. All rights reserved.

https://emcrit.org/pulmcrit/racism-oximetry/ (Dec 2020)

PulmCrit – Dismantling the systemic racism of pulse oximetry

December 21, 2020 by Josh Farkas – 9 Comments



Oximetry is fundamental to critical care. Consequently, even small biases in pulse oximetry measurements could have real clinical impact (especially when leveraged across innumerable measurements among thousands of patients).

Racial bias in pulse oximetry was the subject of two studies in 2005 and 2007. The topic was then largely ignored over the past 13 years. I was vaguely aware of this issue, but given a lack of modern research on it I had assumed that it was a technical glitch restricted to older pulse oximeters. A fresh publication in the New England Journal of Medicine shows that I was wrong. Let's start by reviewing the evidence.

https://doi.org/10.1097/00000542-200504000-00004 (April 2005)

CLINICAL INVESTIGATIONS

Anesthesiology 2005; 102:715-9

* 2005 American Society of Anesthesiologists. Inc. Lippincott Williams & Williams. Inc.

Effects of Skin Pigmentation on Pulse Oximeter Accuracy at Low Saturation

Philip E. Bickler, M.D., Ph.D., John R. Feiner, M.D., † John W. Severinghaus, M.D.

Background: It is uncertain whether skin pigmentation affects pulse oximeter accuracy at low IIbO₂ saturation.

Methods: The accuracy of finger pulse oximeters during stable, plateau levels of arterial oxygen saturation (Sao,) between 60 and 100% were evaluated in 11 subjects with darkly pigmented skin and in 10 with light skin pigmentation. Oximeters tested were the Nellcor N-595 with the OxiMax-A probe (Nellcor Inc., Pleasanton, CA), the Novametrix 513 (Novametrix Inc., Wallingford, CT), and the Nonin Onyx (Nonin Inc., Plymouth, MN). Semisupine subjects breathed air-nitrogen-carbon dioxide mixtures through a mouthpiece. A computer used end-tidal oxygen and carbon dioxide concentrations determined by mass spectrometry to estimate breath-by-breath Sao2, from which an operator adjusted inspired gas to rapidly achieve 2- to 3-min stable plateaus of desaturation. Comparisons of oxygen saturation measured by pulse oximetry (Spo.) with Sao, (by Radiometer OSM3) were used in a multivariate model to determine the interrelation between saturation, skin pigmentation, and oximeter blas (Spo2 - Sao2).

Results: At 60-70% Sao₂, Spo₁ (mean of three oximeters) overestimated Sao₂ (bias ± SD) by 3.56 ± 2.45% (n = 29) in darkly pigmented subjects, compared with 0.37 ± 3.20% (n = 58) in lightly pigmented subjects (P < 0.0001). The SD of bias was not greater with dark than light skin. The dark-light skin differences at 60-70% Sao₂ were 2.35% (Nonin), 3.38% (Novametrix), and 4.30% (Nellcor). Skin pigment-related differences were significant with Nonin below 70% Sao₂, with Novametrix below 90%, and with Nellcor at all ranges. Pigmentrelated bias increased approximately in proportion to desaturation.

Conclusions: The three tested pulse oximeters overestimated arterial oxygen saturation during hypoxia in dark-skinned individuals.

PULSE oximetry theoretically can compute arterial hemoglobin oxygen saturation from the ratio of the pulsatile to the total transmitted red light divided by the same ratio for infrared light transilluminating a finger, ear, or other tissue. The derived saturation should be independent of skin pigmentation, and many other variables, such as hemoglobin concentration, nail polish, dirt, and jaundice. Several large controlled studies comparing black and white patients (380 subjects)^{1,2} reported no

Received from the Department of Anesthesis and Perioperative Care. University of California at Sam Francisco, San Francisco, California. Submitted for publication July 6, 2001. Accepted for publication November 10, 2004. Supported by a LCSF Department of Anesthesia research fram generated from the clinical testing of pulse oximerers. No memulacturers were involved in any part of the concept, design or analysis of this shuly.

Address reprint requests to Dr. Bichler: Sciences 255, Box 0542, University of California Medical Center, 513, Parassus Avenue, San Francisco, California 94145042, Address electronic mail to bicklerpøfanesthesia.scof.edu Individual article reprints may be purchased through the Journal Web site, www.auesbesiology.org. significant pigment-related errors in pulse oximeters at normal saturation.

However, Severinghaus and Kelleher³ reviewed data from several investigators who had reported anecdotal errors (+3 to +5%) in black patients. 4-7 Model simulations of errors due to various pigments were reviewed by Ralston et al.8 Cote et al.9 reported that nail polish and ink on skin surface can cause errors, a finding confirmed anecdotally by others from fingerprinting ink,¹⁰ henna,11 and meconium.12 Intravenously injected dyes cause transient errors.15 Lee et al.14 found overestimation of saturation, especially at low saturation in pigmented patients (Indian, Malay vs. Chinese). The Technology Subcommittee of the Working Group on Critical Care, Ontario Ministry of Health,15 reported unacceptable errors in pulse oximetry at low saturation in pigmented subjects. Zeballos and Weisman¹⁶ compared the accuracy of the Hewlett-Packard (Sunnvvale, CA) ear oximeter and the Biox II pulse oximeter (Ohmeda, Andover, MA) in 33 young black men exercising at three different simulated altitudes. At an altitude of 4,000 m, where arterial oxygen saturation (Sao₂) ranged from 75 to 84%, the Hewlett-Packard underestimated Sao, by 4.8 ± 1.6%, whereas the Biox overestimated Sao, by $9.8 \pm 1.8\%$ (n = 22). It was stated that these errors, previously reported in whites, were both exaggerated in blacks.

During our many years of testing pulse oximeter accuracy at oxygen saturations as low as 50%, we have occasionally noted unusually high positive bias, particularly at very low saturation levels, in some but not in other deeply pigmented subjects. This investigation was therefore specifically designed to determine whether errors at low Sao₂ correlate with skin color.

All pulse oximeters marketed in the United States are required by the US Food and Drug Administration to have been tested and to be certified as accurate to less than $\pm 3\%$ root mean square error at Sao₂ values between 70 and 100%. The great majority of calibration and confirmation tests have been conducted in volunteer subjects with light skin pigmentation.

The Food and Drug Administration has recently suggested that studies of pulse oximeter accuracy submitted for Food and Drug Administration device approval include subjects with a range of skin pigmentations, although no quantitative requirement has been distributed. We are aware of no data that support this action. If there is a significant and reproducible positive bias at low saturation in dark-skinned subjects, inclusion of

Anesthesiology, V 102, No 4, Apr 2005.

715

^{*} Professor, † Associate Professor, ‡ Professor Emeritus

https://doi.org/10.1097/00000542-199206000-00024 (June 1992)

REVIEW ARTICLE Julien F. Biebuyck, M.B., D.Phil., Editor

Anesthesiology 76:1018–1038, 1992

Recent Developments in Pulse Oximetry

John W. Severinghaus, M.D.,* Joseph F. Kelleher, M.D., M.A.†

CONTENTS

Introduction Methodological Developments Uses of Pulse oximetry The Detection of Hypoxia Monitoring Circulation Role in Preventing Retinopathy of Prematurity Investigative Uses Limitations of Pulse Oximetry Incidence of Failure Low Signal-to-noise Ratio **Probe Position** Vasoconstrictors Low-perfusion Limits Motion Artifact Abnormal Pulses Ventilator-induced Pulse Interference Response Times Ambient Light Electrocautery Interference of Magnetic Resonance Imaging Alternative Sites Reflectance Operation Adults Fetal Skin Pigments, Dyes, Nail Polish Carboxyhemoglobin and Methemoglobin Potential Dangers False Alarms and False Nonalarms Accuracy Effect of Anemia on Errors In Vitro Methods of Testing the Accuracy of Pulse Oximeters **Evaluations of Pulse Oximeter Performance and Features** Does Pulse Oximetry Increase Patient Safety? Effect of Pulse Oximetry on Other Monitoring Methods Regulation, Insurance, Standards, and Cost Possible Effects of Oximetry on Anesthesiologists

* Department of Anesthesia and Cardiovascular Research Institute, University of California, San Francisco, California.

† Anesthesia Service Medical Group, Inc., San Diego, California. Received from the Department of Anesthesia and Cardiovascular Research Institute, University of California, San Francisco and the Anesthesia Service Medical Group, Inc., San Diego, California. Accepted for publication January 21, 1992.

This review contains portions of Severinghaus JW: Oximetry: Uses and limitations. ASA Refresher Courses in Anesthesiology 19:139-152, 1991. Pulse oximeters now occupy most critical care arenas and virtually every operating room in the United States. They are manufactured by more than 35 firms, with 1989 annual world wide sales estimated at 65,000 units valued at \$200 million.

In January 1989, two comprehensive reviews of pulse oximetry were published. One gave relative emphasis to theory of operation and other technical aspects,¹ while the other focused primarily on clinical issues.² Since the reviews of 1989 were completed, more than 500 additional publications have described methods, uses, problems, progress, and effects of pulse oximetry—135 of them in the 6-month period prior to the end date of this review (October 1, 1991). The past 3 yr have seen a variety of other reviews concerning some recent developments³⁻¹² as well as the history of pulse oximetry.^{13,14}

The purpose of this article is to summarize the literature on pulse oximetry that has appeared since the major reviews of early 1989. Expressions such as "before 1988" and "since 1988," unless otherwise indicated, refer herein to the mid-1988 cutoff date for the references appearing in those reviews.

Methodological Developments

There is relatively little to report as to methodological advance in pulse oximetry since 1988. A potential exception is surface reflectance ("surface") oximetry, which has received significant recent experimental attention but does not appear ready for widespread clinical use. In 1988, a review of pulse oximetry² could dismiss the topic of reflectance oximetry by observing that it was "one desirable possibility for the future of pulse oximetry . . . in which measurements of reflected light would allow monitoring at nontransilluminable sites, such as the fetal presenting part during labor." With the possible exception of its use in labor, reflectance oximetry is still largely investigational. Some models now couple the oximeter to the ECG (see "Limitations of Pulse Oximetry: Motion

Address reprint requests to Dr. Severinghaus: 1386HSE, UCSF, San Francisco, California 94143.

Key words: Blood: oxygen measurement; oxygen saturation. Complications: hypoxia; postoperative hypoxia; respiratory problems. Equipment: optical; oximeters. Monitoring: oximetry.

https://doi.org/10.1378/chest.97.6.1420 (June 1990)

Reliability of Pulse Oximetry in Titrating Supplemental Oxygen Therapy in Ventilator-Dependent Patients*

Amal Jubran, M.D.; † and Martin J. Tobin, M.D., F.C.C.P.‡

Pulse oximetry is widely used in the critical care setting, but few studies have examined its usefulness in clinical decision making. One area where pulse oximetry might be useful is in the titration of fractional inspired O, concentration (FIo,) in ventilator-dependent patients. Unfortunately, documented guidelines for this use do not exist, and in a survey of directors of intensive care units, we found that they employed a wide range of target O, saturation (SpO,) values. Consequently, we undertook a study to determine if SpO, could be reliably substituted for measurements of arterial O, tension (PaO,) when adjusting FIO, in ventilatordependent patients. We examined a number of SpO, target values in 54 critically ill patients aiming for a PaO, of ≥60 mm Hg, while minimizing the risk of O₄ toxicity. In white patients, we found that a SpO₁ target of 92 percent was reliable in predicting a satisfactory level of oxygenation.

However, in black patients, such a SpO₁ reading was commonly associated with significant hypoxemia (PaO, as low as 49 mm Hg), and a higher SpO, target, 95 percent, was required. In addition, inaccurate oximetry readings (ie, >4 percent difference between SpO, and direct SaO, measurements) were more common in black (27 percent) than in white patients (11 percent, p<0.05). In conclusion, a SpO, target of 92 percent was reliable when titrating supplemental O, in white patients receiving mechanical ventilation; however, in black patients, such a SpO, reading was commonly associated with significant hypoxemia, and a higher SpO, target, 95 percent, was required to ensure a satisfactory level of oxygenation. (Chest 1990; 97:1420-25)

drawn. While she had a pulse oximeter probe attached to her finger, we did not know what optimal O₂

SpO, = pulse oximetry based oxygen saturation

Pulse oximetry has gained enormous popularity since its introduction into clinical medicine. In 1988, there were 23 manufacturers of pulse oximeters, 45 different models available, and approximately 45,000 units in use.1 Indeed, Severinghaus and Astrup² recently considered pulse oximetry to be "the most significant technologic advance ever made in monitoring the well-being and safety of patients during anesthesia, recovery, and critical care." Although pulse oximeters have become ubiquitous in intensive care units, their role in clinical decision making is not well defined because most studies have focused on their accuracy rather than on their clinical usefulness.³

One area where oximetry might be useful is in the titration of optimal levels of supplemental oxygen (O2). Recently, we had to totally depend on pulse oximetry to adjust O₂ therapy in a patient with severe adult respiratory distress syndrome who had significant hypoxemia (arterial O₂ tension of 55 mm Hg) while receiving a fractional inspired O₂ concentration (FIO₂) of 0.80. This young woman refused permission for the insertion of an arterial catheter, and also requested that no further arterial blood gas samples would be

*From the Division of Pulmonary Medicine and Critical Care, University of Texas Health Science Center, Houston. †Research Fellow. ‡Associate Professor

Manuscript received September 21; revision accepted December

Reprint requests: Dr. Jubran, University of Texas at Houston, 6431 Fannin, Houston 77030

1420

saturation target was to aim for (ie, ensure adequate oxygenation while minimizing the risk of O₂ toxicity), because published guidelines do not exist. Consequently, we undertook a systematic study of this problem consisting of two parts. First, we conducted

METHODS

A telephone survey was conducted to determine if pulse oximetry is being used to evaluate the response to titrations in Flo.. The medical directors of 25 hospitals (both university affiliated and private community hospitals) from various geographic locations throughout the United States were contacted. The medical director was asked whether he/she used pulse oximetry to titrate FIo, in ventilator-dependent patients, and, if so, what target SpO, he/she

Patient Study

Ratients: Fifty-four patients admitted to the medical intensive care units of Hermann Hospital and M. D. Anderson Hospital in

Reliability of Pulse Oximetry in Titrating Supplemental Oxygen (Jubran, Tobin)

a telephone survey of directors of intensive care units (ICUs) throughout the United States to determine if pulse oximetry is being employed to titrate FIO₂ in ventilator-dependent patients and the manner in which this is done. Secondly, we prospectively examined a number of algorithms to determine if pulse oximetry is reliable in assessing alterations in oxygenation that result from titration of FIO₀ in ventilatordependent patients.

Hospital Survey

employed.

https://pubmed.ncbi.nlm.nih.gov/2463349/ (Fall 1987)

Pub Med.gov	Search PubMed
	Advanced
	Save
> J Perinatol. Fall 1987;7(4):329	9-30.
Skin pigmentatio pulse oximetry	on as an influence on the accuracy o
J R Emery ¹	
Affiliations + expand PMID: 2463349	
No abstract available	
Similar articles	
The accuracy of pulse oxime Anderson JV.	etry in neonates: effects of fetal hemoglobin and bilirubin.
J Perinatol. 1987 Fall;7(4):323, PMID: 2463346 No abstract ava	ailable.
Physiology of oxygenation a	nd its relation to pulse oximetry in neonates.
Hay WW Jr.	
Hay WW Jr. J Perinatol. 1987 Fall;7(4):309-19.	

Datasheets for Datasets by Timnit Gebru et al.

Example datasheet for Pang and Lee's polarity dataset [ACL 2004]

Thumbs Up? Sentiment Classification using Machine Learning Techniques

Motivation For what purpose was the dataset created? Was from a specific built in mind? Was there a specific gap that reached to be file d? Please provide

Movie Review Polarity

The dataset was created to enable research on predicting sentiment polarity: given a pie ce of English lest, predict whether it has a positive or negative affect-or stance-toward its topic. It was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.

Who created this dataset (e.g., which eram, research group) and on behalf of which entity (e.g., company, institution, organic adon)? The dataset was created by Bo Pang and Lillian Lee at Cornell

Who lunded the ontation of the dataset? If there is an associated grant, please provide the name of the granter and the grant mame and number. unding was provided though five distinct sources: the National Science Roundation, the Department of the Interior, the National Business Center Cornell University and the Stean Ex-Any other commones?

Composition

stamps is shown in Flaure 1

What do the insumces that comprise the dataset represent (e.g., doc-umantes, phones, people, cotonitical? An same multiple types of lo-stances (e.g., molver, users, and natings; people and interactions is tensor them; nocks and adge al? Please provide a discription. The instances are movie reviews extracted from newsgroup post ment rating for whether the text come sponds to a review with a rating that is either strongly positive high number of stars) or strongly regative (low number of stars).

The polarity rating is binary {positive, negative}. An example in-How many instances are there in total (of each type, if appropriate ?? There are 1400 instances in total in the original (v1.x versions) and 2000 instances in total in v2.0 (from 2014).

Does the dealers contain all possible instances or is it a sample (not necessarily random) of instances from a larger sof? If the dataset is a sample, then what is the larger at? Is the sample representative of the larger and (e.g., gaugesphic coverage)? If so, planae describe how this representativeness was validated writed. If it is not specaritative of the larger as t, bissae describe why not leg., to cover a more deverae nonge of instances, because instances were withheid or unavailable.) The dataset is a sample of instances. It is (presumably) intended to be a random sample of instances of movie revies from newsgroup postings. No tests were run to determine representative-

¹Internation in this database is taken frees one of the warraw, any errors the wear interdanced on our fault. http://www.ca.com/likeds/prophilphof/ novie-mvive-data/inter/interdata.interdata/interdata/interdata/ com/likeds/prophilphof/interdata/interdata/interdata/interdata/ com/likeds/prophilphof/interdata/interdata/interdata/interdata/ list. http://www.ca.com/likeds/interdata/int

these are words that could be used to describe the emotions of joins sayles" raciers in his laient , limbo , but no , i use them to describe myself all Consider as no native, hence, tak no, Rigant I. An example "negative polarity" instance, taken from the El

What data does each insumer consists of? "Dae" data is g, unpro-memod text or imaginalor instants? In other case, pixee provide a de-oricide.

Each instance consists of the text associated with the review, with invices ratines information removed from that test (some errors vere found and alter fixed). The text was down-cased and HTMI tags were removed. Boilemiaie new saroup header footer te xi was removed. Some additional unspecified automatic fillering was done. Each instance also has an associated target value: a pos-tithe (+1) or negative (-1) rating based on the number of stars that that review gave (idealis on the mapping from number of stars to polarity is given below in "Data Preprocessing"). is them a label or target associated with each instance? If so, please provide a description.

Is any information missing from individual instances? If as, please provide a deactplion, explaining why this information is missing (s.g., be-rates it was unavailable). This does not include intertionally removed formation, but might include, e.g., reducted text Every thing is included. No data is missing. Are relationships between individual instances made explicit (e.g.

poster name and email address, so some information could be acted if needed

Are there accommended data splits to g., unining, develop-monvalidation, ustiling?? If we please provide a description of these splits, explaining the reduced tenth of them. The instances come with a "cross-validation tag" to enable replication of cross-validation experiments; results are measured in

Are there ary errors, sources of noise, or redundancies in the dataset? If so, pinase provide a description

is the dataset soli-contains d, or does it link to or otherwise roly or Is the calculated solution of course in time to or contention tray on contential reconcerned to a solution of the solution of the solution to or solits on external measures, a) are then guarantees that they will exist, and means contraited, we then the solution of the complete defauent time. It is a solution to contrait, a set them to a solution of the complete defauent time, because a solution of the defauence is a solution of the solution of the contrait measures as they as label at the time the dataset that contrait, of the mean and measures that they are an an an and the solution of the contrait measures that might capty to a that our our? Please provide the cologities of all external solutions of the contraint of the solution of all external

execution and any matrictions associated with them, as well as links or other access points, as appropriate. Does the datage i contain data that might be considered confidentia le.e., data that is provide d by is gai privile as or by decernation; con

Thumbs Up? Sentiment Classification using Machine Learning Technique

Schneally, data that includes the content of individuals non-public Doos the dataset contrain data that if viewed directly might be offer Some movie reviews might contain moderately inappropriate or effensive language, but we do not expect this to be the norm.

Does the dataset relate to people? If not you may also the remaining

Movie Review Polarity

Does the dearse identify any subpopulations (e.g., by age, gender)? If so, please dears has how these subpopulations are identified and provide a dearsign of their a specifies distributions within the dataset.

is it possible to identify individuals (i.e., one or more neural per-sons), either directly or indirectly (i.e., in combination with other data) from the dataset? If as, please directly how Does the dataset constitut data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orienus belie is, political opinions or union mahine ne iscasions; Rinancial or health data; biometric or genetic data; forms of governmen identification; such as social acutiny numbers; criminal ony)? If so, please provide a description The raw form of the dataset contains names and email addresses,

out these are already public on the internet newsaroup. Any other commones?

Collection Process

users' novie raings, social movers links?" If an piece describe has fease elationing an made actict Nore explicitly, though the original newsgroup postings include

dataset have control over usage of their data or the ability to re-move their information from the dataset entirely. How was the data associated with each instance acquired? Was the

The data was mostly observable as raw text, except the la hels were extracted by the process described below. The data was collected by downloading reviews from the IMDb archiv of the rec. arts. novies. reviews newsgroup, at hep

makes a loost room The insta

What mechanisms or procedures were used to collect the data lo.g., hardwate apparatus or sensor, manual human curation, solwate pro-gram, softwate AP07 Haw were the ac mechanisms or procedure a valgram, sofeware AP07 Unknym

If the dataset is a sample from a larger set, what was the sam-pling strategy (e.g., deterministic, probabilistic with specific sam-

The sample of instances collected is English movie reviews from the rec.arts.movies.reviews newsgroup, from which a "number of stars" rating could be extracted. The sample is limited to forty reviews rer unique author in order to achieve broader overage by authorship. Beyond that, the sample is arbitrary Who was involved in the data collection process (e.g., sudones, crowdworkers, conversions) and how were they compensated (e.g., how much were crowdworkers paid)? Over what empirants was the data collected? Dury this firminants

match the creation time have of the data searcheed with the instances is g, recent creat of old news articles? If not please describe for time have in which the data sourceast with the instances was created Unknown Were any orbital review processes conduced (e.g., by an inselu-denail twice board)? I so, please provide a description of these review processes, including the outcomes, so well as a link or other second point

to any autoorting documentation Does the dataset relate to people? If not, you may skip the remaining The dataset relates to people in that the reviews the methes are authored by people. Personally identifying information (e.g., email

Did you colle to the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? The data was collected from newsaroups.

Similar to Composition, this section should be read during the Were the individuals in question notified about the data collection? Final please describe (or show with accentrates or other information) have notifier was provided, and provide a link or other accente points to or other-wise reproduce, the exact language of the notification itself. initial planning phase, and filled out during the collection of data. Again, there questions provide general transparency into the makeup of the data help both the dataset creator and dataset No. The data was crawled from mublic web sources and the auconsumer uncover risks and potential harms, for example by thors of the posts presumably knew that their posts would be pubquestioning whether those whose information is contained in the lic, but there was no explicit informing of these authors that their

posts were to be used in this way. Did the individuals in question conserve to the collection and use of their data? If we, pieze describe (or show with conservative or other information) have common wave magnetical and provided, and previde a link or other accents point is, or otherwise reproduces, the work language is which the information.

the density descended (e.g., nor that, model and gal), in ported by ad-plants in g, servey responses), or indirectly internet/derived incertains (e.g., part-of-speech tags, model/sawid generations for age or the regargit)? If data was reported by addictor or indirectly internet/derived from other data, was the data validation's write off 10 ag plasme describe how. No (see previous que stion).

If consome was obtained, were the consoming individuals provided with a mechanism to revolue shale consom in the learne or for contain 1885 ? If its, please provide a classificition, as well as a link or other access point to the mechanism (if appropriate). Has an analysis of the potential impact of the dataset and its use on data subjects to p, a data protection impact analysis/been con-

BECREO7 II so, please provide a cleaription of this analysis, including the outcomes, as well as a link or other access point to any supporting docu-

A my helper homenous?

Preprocessing/cleaning/labeling Was any proprocessing chaning labeling of the data done is.g., dis-creditation of bactering, extendations, part-of-speech tagging, SIFT labeling caracterion, emviol of historators, processing of missing val-ues of the spectra provide a description. If not, you may align the remain-der of the spectra in in the accident. Instances for which an explicit rating could not be found were discarded. Also only instances with strongly position or strongly-negative ratings were retained. Star ratings were extracted by automatically looking for text like "++++ out, of" in the review, usine that as a label, and then removing the corresponding text. When the star rating was out of five stars anything at least four was considered positive and anything at most two negative; when out of four, three and up is considered positive, and one or less is considered negative. Occasionally half stars are missed which affects the labeling of negative examples. IDCR/S Everything in the middle was discarded. In order to ensure that sufficiently many authors are represented, at most 20 reviews (per positive/negative label) per author are included. In a later version of the dataset (v 1.1), non-English reviews were also removed. Some preprocessing errors were caught in later versions. The fol-

Movie Review Polarity

lowing fixes were made: (1) Some reviews had rating information in several places that was missed by the initial filters; there are removed. (2) Some reviews had ure specied/unsarged nances and these were fixed. (3) Sometimes the bollerplate removal removed too much of the text. Was the "raw" data saved in addition to the propro-cessed cleare distributed data (e.g., to support unamicipated Autor uses)? If so, please provide a link or other scenes point to the

Yes. The dataset itself contains all the raw dta. is the software used to preprocess-clean/label the insurces avail-able? If as please provide a link or other access point.

Any other commones?

Uses

Has the dataset been used for any usits already? If so, please provide At the time of publication, only the original paper http://cockie/l. gol/ed/scionoscovi. Between then and 2012, a collection of papers that used this dataset was maintained at http://www.ca.com/ll

a %2Drawlerer%2 Delated is there a repository that links to any or all papers or systems that

to the a separately manufactor any or an paper of synahls that use the databil? If so, please provide a link or other access point There is a repository, maintained by PangLee through April 2012, al. http://www.cu.com/lindu/proje/pabo/revin%20Devine%

Thumbs Up? Sentiment Classification using Machine Learning Techniques

What induct satis could the dataset he used inc? The dataset could be used for anything related to modeling or understanding movie reviews. For instance, one may induce a lexicon of words/phrates that are highly indicative of sentiment polarity, or learn to automatically generate movie reviews.

is there anything about the composition of the dataset or the way it was colloced and preprocessed/cleare.dtaboled that might impact R000 02257 For example, is there anything that a future user might need to know to avoid uses that could would in unfair to atment of individuals o In where to avoid using that could wait in urnar to attend or nowidual or groups (s.g., stanodyping quality of anivitas bases) or other undealedain harma (s.g., financial harma, legal riska) II so, please provide a descrip-for, is then anything a threat user could do be militate these undealedain

There is minimal risk for harm: the data was already public, and in the preprocessed version, names and email addresses were re-

Are show useks for which the dataset should not be used? If any please provide a description. provide a description. This data is collected solely in the movie review domain, so systems trained on it may or may not generalize to other writt. ent mediction tasks. Come quently, such assigns should not

without additional verification be used to make consequential decisions about receive Any other comments?

Distribution Will the dataset be distributed to third panies outside of the en-try (e.g., company, institution, organization) on behalf of which the dataset was (702000? If an please provide a direction). Yes, the dataset is publicly available on the internet

How will the dataset will be distributed to g., tarball on websiat, API, CitHubl? Done the dataset have a digital object identifier (DOI)? The dataset is distributed on Bo Pang's webpage at Cornell: http: //www.cs.cornell.edu/prople/pabo/movie-nvine-data. The dataset does not have a DOI and there is no redundant archive.

When will the dataset he distributed?

Will the datase to distributed under a copyright or other intellectual roperty (IP) license, and/or under applicable terms of use (ToU)?

so, plusan dearthe this lannar and/or To(1, and provide a link or other accurs point to, or otherwise superdates, any solvent literating terms or To(1, as well as any lines associated with these restrictions. The crawled data copyright he longs to the authors of the reviews unless otherwise stated. There is no license, but there is a request to cite the corresponding paper if the dataset is used: Thanks ap? Sensimene classification using machine learning sechniques. Bo Pane, Lillian Lee, and Shivakumar Vaithvanathan. Proceedings of EMNLP 2002

the reconcilments, html

Unknown Any other comments

creator think through their plans for updating, adding to, or fixing errors in the dataset, and expose these plans to dataset consum-

How can she owner/curatormanager of the datasets be conserved is g., email address/?

dates and known errors are specified in higher version parameters and diff files. There are several version s of these vilo

2Ddets/READVE.1.1 and http://www.ca.comelle.du/people/pate maxis-awine-data/dilitet; w2.0: http://www.ca.comeil.edu/people/pab movie %2Deview %2Ddeter/coldete, README 2.0.61. Updates are listed on the dataset web page. (This datasheet largely summarizes

Will the dataset be updated is.g., to correct labeling errors, add new instances, deb to instances?? If as, please deacribe has often, by when, and how updates will be communicated to users is g, m CMLAP2

This will be posted on the dataset we brase

If the datazet i to lates to propile, and there applicable limits on the in-tention of the data associated with the instances to g, were individu-als in quastion told that short data would be mained for a fixed period of time and then deliced/? If so, please do actin these limits and explain how they will be a microad

The dataset has already been updated; older versions are kept around for consistency.

https://arxiv.org/abs/1803.09010 https://cacm.acm.org/magazines/2021/12/256932-datasheets-for-datasets/ (Dec 2021)

Movie Review Polarity Thumbs Up? Sentiment Classification using Machine Learning Technique Have any third panias imposed IP-based or other reservicions on the data associated with the instance S7 if nc, picase describe frees watch-tions, with provide a link or other accuracy point is, or otherwise reproduce, way relevant beaming terms, as well as any their same stated with them

I obtars ware to are and suggestimabelid on contributes to the second is shown a machinism to them to do so? If its, please provide a disordy-for. Will frame combulence to validand verified? If its please do software has. If not they not it is then a provide a share/plate. The second second second second second second second second second to second authors although the second incorporating fixes/extensions

Do any exponencempois or other regulatory restrictions apply to the dataset or to individual insurances? Even change datasets from matrice Any other commons? form, and provide a link or other access any supporting documentation.

Maintenance

This section should be completed once the dataset has been con-structed, before it is distributed. These questions help the dataset Bo Pang is supporting/maintaining the dataset.

Unknown is there an erratum? If an ok any provide a lok or other appears point. Since its initial release (v0.9) there have been three later releases (v1.0, v1.1 and v2.0). There is not an explicit erratum, but up-

http://www.ca.comeil.edu/people/pabs/movie-envire-data/README; ×1.1http://www.ca.comeil.edu/ce/ce/a/ce/co/wow/a%2Dev/aw%

these sources)

Will older versions of the dataset continue to be sup-period/hose-dimainatined? If an please describe how. If not, please describe how its obseivations will be communicated to same.

Datasheets for Datasets by Timnit Gebru et al.

Example datasheet for Pang and Lee's polarity dataset [ACL 2004]

Thumbs Up? Sentiment Classification using Machine Learning Techniques

The dataset was created to enable research on predicting sentiment polarity: given a piece of English lext, predict whether it has a positive or negative affect-or stance-toward its topic. It was created intentionally with that task in mind, focusing on movie reviews as a place where affect/sentiment is frequently expressed.

Who created this dataset (e.g., which eram, research group) and on behalf of which entity (e.g., company, institution, organic adon)? The dataset was created by Bo Pang and Lillian Lee at Cornell

Who lunded the ontation of the dataset? If there is an associated grant, please provide the name of the granter and the grant mame and number. unding was provided though five distinct sources: the National Science Foundation, the Department of the Interior, the National Business Center, Cornell University, and the Stoan Fo Low odpor commonly

Composition

The instances are movie are lews extracted from newseroup post ngs, together with a sentiment rating for whether the text corresponds to a review with a rating that is either strongly positive (high number of stars) or strongly regative (low number of stars). The polarity rating is binary {positive, negative}. An example instance is shown in Figure 1.

How many instances are there in total (of each type, if appropriate)? There are 1400 instances in total in the original (v1.x versions) and 2000 instances in total in v2.0 (from 2014).

Does the dealers contain all possible instances or is it a sample (not necessarily random) of instances from a larger sof? If the dataset is a sample, then what is the larger at? Is the sample representative of the larger and (e.g., gaugesphic coverage)? If so, planae describe how this representativeness was validated writed. If it is not specaritative of the larger as t, bissae describe why not leg., to cover a more deverae nonge of instances, because instances were withheid or unavailable.) The dataset is a sample of instances. It is (presumably) intended to be a random sample of instances of movie revies from news-

group postings. No tests were run to determine representativeternation in this databast is taken from one of the scarces; any corres on instalant are not fast. http://www.co.com/al.ad.group/is/past/ writes-data; // http://co.last.gov/file/dov/adata/file/data.file/ADME-1. http://www.co.com/al.ad/mode/data/file/data.file/ADME-1. http://www.co.com/al.ad/mode/data/file/data.file/ADME-1.

Motivation these are words that could be used to describe the emotions of joins sayles" machers in his baixed , limbo , but no , i use them to describe myself alle conserve in its target, metho, but no , into the two describe negative functions that the first high temperature is index generated. It can be give its provide the two describes the second s Rgant I. An example "augative polarity" instance, taken from the Eli

What data does each insumer consists of? "Dae" data is g, unpro-memod text or imaginalor instants? In other case, pixee provide a de-oricide.

Each instance consists of the text associated with the review, with invices ratines information removed from that test (some errors vere found and alter fixed). The text was down-cased and HTMI tags were removed. Boilemiaie new saroup header footer te xi was removed. Some additional unspecified automatic filterine was done. Each instance also has an associated target value: a pos-tithe (+1) or negative (-1) rating hased on the number of stars that that review gave (details on the mapping from number of stars to polarity is given below in "Data Preprocessing").

is shere a label or sarges associated with each insuance? If so, please provide a descriteion is any information missing from individual insumo 57 if so, picase provide a directption, explaining why this information is missing (e.g., to-pause it was unwailable). This does not include intertionally removed

information, but might include, e.g., reducted text. Every thing is included. No data is missing. Are relationships between individual instances made explicit (e.g., user: invite raining, social mercels links?) H se please describe fee these describes the describes the explicitly, though the original newsgroup posings include

poster name and email address, so some information could be acted if needed Are there accommended data splits to g., unining, develop-monvalidation, ustiling?? If we please provide a description of these splits, explaining the reduced tenth of them. The instances come with a "cross-validation tag" to enable repli-

cation of cross-validation experiments; results are measured in Are there ary errors, sources of noise, or redundancies in the

dataset? If so, pinase provide a description is the dataset soli-contains d, or does it link to or otherwise roly or Is the calculated solution of course in time to or contention tray on contential reconcerned to a solution of the solution of the solution to or solits on external measures, a) are then guarantees that they will exist, and means contraited, we then the solution of the complete defauent time. It is a solution to contrait, a set them to a solution of the complete defauent time, because a solution of the defauence is a solution of the solution of the contrait measures as they as label at the time the dataset that contrait, of the mean and measures that they are an an an and the solution of the contrait measures that might capty to a that our our? Please provide the cologities of all external solutions of the contraint of the solution of all external

resources and any matrictions associated with them, as well as links or other access points, as appropriate.

Does the datage i contain data that might be considered confidentia le.e., data that is provide d by is gai privile as or by decernation; con

Thumbs Up? Sentiment Classification using Ma-

Schneially, data that includes the content of individuals non-public ommunications/7 IF so, please provide a description. Doos the dataset control of the share if viewed disords might be offer dea, insuléng, threasoning, or might otherwise cause andery ? If so, dease describe why Some movie reviews might contain moderately inappropriate or Who was involved in the data collection process (e.g., sudones, crowdworkers, conversions) and how were they compensated (e.g., how much were crowdworkers paid)? offensive language, but we do not expect this to he the norm.

Does the detected to leave to people? If not you may also the remaining manufactor in this section. Over what simelrame was the data collected? Dury this firminant Does the dataset identify any subpopulations (e.g., by age, gender) If an pieces describe the free subpopulations are identified and provid a description of their respective databations within the dataset. match the creation time have of the data searcheed with the instances is g, recent creat of old news articles? If not please describe for time have in which the data sourceast with the instances was created

is it possible to identify individuals (i.e., one or more neural per-sons), either directly or indirectly (i.e., in combination with other data) from the dataset? If as, please directly how Unknown

Movie Review Polarity

Any other commones?

strings inch tors Decises

Unknewn

Were any ordical review processes conducerd (e.g., by an insels-sonal toylew board?) I so please provide a description of these review processes, including the outcomes, as well as a link or other access pole Daes the dataset constitut data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orien-tations, reliaious belie is, political coinfors or union mamberships, or iccasions; Francia or health das; biometric or genetic das; forms of governmen identification; such as social acutiny numbers; criminal DIV/? If so, please provide a description

hels were extracted by the process described below. The data

was collected by downloading reviews from the IMDb archive

of the rec. arts. novies. reviews newsgroup, at hep

What mechanisms or procedures were used to collice the data is .g., hardware apparatus of sonsor, manual human curation, solware pro-gram, solware APII? How were these mechanians or procedure val-daw??

The dataset relates to people in that the reviews the mselves are au-The raw form of the dataset contains names and email addresses, but these are already public on the internet newsgroup. thored by people. Personally identifying information (e.g., email was removed.

to any autoorting documentation

Did you colle to the data from the individuals in question directly, or characteristic to the parties or other sources is .g., websites??

Does the dataset relate to people? If not, you may skip the minaking



the makeup of the data help both the dataset creator and dataset No. The data was crawled from mehiir web sources and the auconsumer uncover risks and potential harms, for example by thors of the posts presumably knew that their posts would be pubquestioning whether those whose information is contained in the lic, but there was no explicit informing of these authors that their dataset have control over usage of their data or the ability to re-move their information from the dataset entirely. How was the data associated with each instance acquired? Was the posts were to be used in this way.

Did she individuals in quession corrects to the collection and use of their data? If as, piezes describe (or show with scenarioto or other information) have consert was anygented and provided, and provide a link or other access point is, or otherwise reproduce, the exact language to which the infolduals conserted. the most is an exactly come to be a substantial state of the second state of the secon No (see previous que stion). The data was mostly observable as raw text, except the la-

If consome was obtained, were the consoming individuals provided with a mechanism to revolue shale consomitin the leaser or for contain 1885 ?? If is, please provide a classificition, as well as a link or other access point to the mechanism (if appropriate).

Has an analysis of the potential impact of the dataset and its use on data subjects to p, a data protection impact analysis/been con-BECREO7 II so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting docu-

If the dataset is a sample from a larger set, what was the sam-pling strategy (e.g., deterministic, probabilistic with specific sam-A my helper homenous?

The sample of instances collected a Preprocessing/cleaning/labeling without the same control is the same c Instances for which an explicit rating could not be found were discarded. Also only instances with strongly-positive or strongly-negative ratings were retained. Star ratings were extracted by automatically looking for text like "++++ out, of "" in the review, usine that as a label, and then removing

the corresponding text. When the star rating was out of five stars, anything at least four was considered positive and anything at most two negative; when out of four, three and up is considered positive, and one or less is considered negative. Occasionally half stars are missed which affects the Tabeling of negative examples. Everything in the middle was discarded. In order to ensure that sufficiently many authors are represented, at most 20 reviews

(per positive/negative label) per author are included. In a later version of the dataset (v 1.1), non-English reviews were also removed.

Some preprocessing errors were caught in later versions. The foltrwing fixes were made: (1) Some reviews had rating information in several places that was missed by the initial filters; there are removed. (2) Some reviews had ure specied/unnarsed ranges and these were fixed. (3) Sometimes the bollerplate removal removed

too much of the text. Was the "raw" data saved in addition to the propro-cessed cleared table bid data (e.g., to support unamicipated know uses? If as pinate provide a link or other scena point to the

Yes. The dataset itself contains all the raw dta. is the software used to preprocess-clean/label the insurces avail-able? If as please provide a link or other access point.



USES^{T so, pleases provide} Ray the datator her At the time of princessor, only an original paper http://oochol. go/pdfea/94090004-1. Between then and 2012, a collection of pa-

pers that used this dataset was maintained at http://www.ca.com/ll N2DreviewN2Ddate the recontinue to html. is there a repository that links to any or all papers or systems that

to make a separatory management of any or an paper's of systems that use the dataset? If so, please provide a link or other access a point. There is a repository, maintained by Pang/Lee through April

2012. al http://www.ca.comelladu/pabe/pabe

Is there anything about the composition of the dataset or the way is was collected and preprocessed cleared labeled that might impact was conserve any population of the acycling final faither and the population of the first BBSF for presentable in the acycling final faither care thigh result to force to avoid some final could musil to orbit to achieve to readinable faither acycling final of any to analy of poly could be accelerated and the state of the first source of the first source of the first source or cold so the first source or cold the first source or the first source or cold the first source or the first There is minimal risk for harm: the data was already public, and in the preprocessed version, names and email addresses were re-IDCR/N

> Are there useds for which the dataset should not be used? If so, pie see provide a description. provide a description. This data is collected solely in the movie review domain, so

systems trained on it may or may not generalize to other writt. nent prediction tasks. Come quently, such systems should not without additional verification-be used to make consequential decisions about receive

Distribution

MEMORY WAS CREATED ? If an official provide a description Yes, the dataset is publicly available on the internet How will the dataset will be distributed is go, tarbail on website, API, Girlikil? Done the dataset have a dipit object kindlar (DOI?) The dataset is distributed on Ro Pang's website dataset (dataset News excornelled) application with writer data. The dataset dataset not have a DOI and there is no redundant archive.

unless otherwise stated. There is no license, but there is a request to cite the corresponding paper if the dataset is used: Thanks ap? Sensimene classification using machine learning sechniques. Bo Pane, Lillian Lee, and Shivakumar Vaithvanathan. Proceedings

of EMNLP 2002

When will the dataset he distributed? The dataset was first released in 2002. Will the datase to distributed under a copyright or other intellectual

we use catable tas unservicedo tantos a copyrigan to estimation manketoso poperer (2011) lettorosa, andro ruto dura applicable arritos of uso (1010)? as, pleasa danazha dia llarana andro Toll, and posoda a la lick or ober-anna polet ay, or oberevien applicable, ay evidente llaraning itema or Toll, as well as any los associated with twas reactediors. The crase ted data copy right tengos to the automos of the reviews

The dataset has already been updated; older versions are kept around for consistency.

Movie Review Polarity

Thumbs Up? Sentiment Classification using Machine Learning Technique

incorporating fixes/extensions

Any other commons?

Is there a machanism for them to do so? If so, please provide a descrip-for. Will found contributions be validated dverified? If so, please the softs has. If not, why not? Is there a process for communicating/databating frame contribution to other users? If so, please provide a theoription.

Others may do so and should contact the original authors about

Have any third panies imposed IP-based or other reserictions on the data associated with the instances? Fac, piezes describe frees with-fam, and provide a link or other access point is, or otherwise reproduce, any relevant forming terms, as seed as any line a succelated with them

Do any exponencempois or other regulatory restrictions apply to the dataset or to individual insurances? Even change datasets from matrice form, and provide a link or other access any supporting documentation. Unknown

Any other comments

Maintenance

creator mink unrough their plans for updating, adding 40, or string errors in the dataset, and expose these plans to dataset consumer Who is supporting hosting maintaining the dataset? Bo Pang is supporting/maintaining the dataset.

How can she owner/curatormanager of the datasets be conserved is g., email address/? Unknown

is there an erratum? If an ok any provide a lok or other appears point. Since its initial release (v0.9) there have been three later releases (v1.0, v1.1 and v2.0). There is not an explicit erratum, but updates and known errors are specified in higher version parameters and diff files. There are several versions of these: v) http://www.ca.comeil.edu/propio/pabo/movio-mvior-data/VEADME; s of these vilo

×1.1http://www.ca.comell.edu/ce.co/e/ce/co/movie%2/Device% 2Ddets/READVE.1.1 and http://www.ca.comelle.du/people/pate movie-envire-data/dif.txt; v2.0: http://www.cs.comeil.edu/people/pabs movie %2Deview %2Ddeter/coldete, README 2.0.61. Updates are listed on the dataset web page. (This datasheet largely summarizes these sources)

Will the dataset be updated is.g., to correct labeling errors, add new instances, debta instances?/ If at, please describe how often, by whom, and how updates will be communicated to users is g., mailing list, CMLA.27

This will be posted on the dataset we brase

If the datazet i to lates to propile, and there applicable limits on the in-tention of the data associated with the instances to g, were individu-als in quastion told that short data would be mained for a fixed period of time and then deleted/? If so, please do actin these limits and explain how they will be a microad

Will older versions of the dataset continue to be sup-period/hose-dimainatined? If an please describe how. If not, please describe how its obsolessment will be communicated to same.

Can we generate "aspirational" data?



Deborah Ramirez
Can we generate "aspirational" data? Aspirational data: synthetic data from ideally fair circumstances



Deborah Ramirez

Rawls' well-ordered society

A society is well-ordered when it ...

"advances the good of its members"

and

"is effectively regulated by a public conception of justice"



John Rawls

A Theory of JUSTICE

> JOHN RAWLS

> > **REVISED EDITION**



John Rawls

Rawls' Principle of Fair Equality of Opportunity (FEO)

- FEO governs the fair allocation of advantageous positions (e.g., high-paying jobs) in society.
- FEO can be formalized as a conditional independence (CI)
 - The probability of securing an advantageous position ought to be independent of protected variables (e.g., race) given justified variables (e.g., talent):

$$Y \perp \!\!\!\perp X_p | X_j$$



RAWLSNET







David Liu

Zohair Shafi

Will Fleisher

Scott Alfeld

<u>Question:</u> Given an unfair (in the Rawlsian sense) outcome and the capability to alter some (but not all) decision-making processes, how can one satisfy FEO?

<u>Answer:</u> **RAWLSNET**, a system for altering the parameters of Bayesian Network (BN) models to satisfy FEO

David Liu, Zohair Shafi, Will Fleisher, et al. RAWLSNET: Altering Bayesian Networks to Encode Rawlsian Fair Equality of Opportunity. *AIES* 2021: 745-755.

RAWLSNET

- RAWLSNET has three components:
 - 1. Learn a BN.
 - 2. Determine relevance to FEO.
 - Update parameters of the BN to satisfy FEO if possible.
 Otherwise, update the parameters to approximately satisfy FEO.

College Admissions

	SES	P	Talent	
	Low	0.8	SES Talent Low	
	High	0.2	High	
			Test	
			High	
			Test High	
			High	
			High	
			College	e
			College	
			Yes	
8	College	P	Yes	
	Low	. 01	Yes	
N	High	0.1		
h	Low	0.25	Job	
h	High	0.95		
-				
	Ta	rget Variab	le Sensitive Variable Justified Varia	ł

High	0.5		
Test	SES	Talent	P
High	Low	Low	0.1
High	Low	High	0.9
High	High	Low	0.25
High	High	High	0.95
College	SES	Test	P
Yes	Low	Low	Variable
Yes	Low	High	Variable
Yes	High	Low	0.1 (Fixed)
Yes	High	High	0.85 (Fixed)

P

0.5

Job	SES	College	P
Yes	Low	Low	0.1
Yes	Low	High	0.9
Yes	High	Low	0.25
Yes	High	High	0.95

Legend

ble

Other Variables

• Admissions policy is θ_{College}

$$\theta_{\text{College}} = \{\theta_{ij} = P(\text{College}|\text{Test} = i, \text{SES} = j)\} \forall i, j \in \{0, 1\}$$



• Admissions policy is θ_{College}

$$\theta_{\text{College}} = \{\theta_{ij} = P(\text{College}|\text{Test} = i, \text{SES} = j)\} \forall i, j \in \{0, 1\}$$

• Optimization problem:

 $\underset{\theta_{\text{College}}}{\operatorname{argmin}} \left[P(\text{Job}|\text{Talent} = 0, \text{SES} = 0) - P(\text{Job}|\text{Talent} = 0, \text{SES} = 1) \right]^2 +$

 $[P(Job|Talent = 1, SES = 0) - P(Job|Talent = 1, SES = 1)]^2$

s.t. $\theta_{i,j} \in \{0,1\} \forall i,j \text{ and } f_{\text{feasibility}}(\theta_{\text{College}}) \leq c_{\text{feasibility}}$



• Admissions policy is θ_{College}

$$\theta_{\text{College}} = \{\theta_{ij} = P(\text{College}|\text{Test} = i, \text{SES} = j)\} \forall i, j \in \{0, 1\}$$

• Optimization problem:

 $\underset{\theta_{\text{College}}{\text{formula}}{P(\text{Job}|\text{Talent} = 0, \text{SES} = 0) - P(\text{Job}|\text{Talent} = 0, \text{SES} = 1)]^2} + [P(\text{Job}|\text{Talent} = 1, \text{SES} = 0) - P(\text{Job}|\text{Talent} = 1, \text{SES} = 1)]^2$ s.t. $\theta_{i,j} \in \{0, 1\} \forall i, j \text{ and } f_{\text{feasibility}}(\theta_{\text{College}}) \leq c_{\text{feasibility}}$



• Admissions policy is θ_{College}

$$\theta_{\text{College}} = \{\theta_{ij} = P(\text{College}|\text{Test} = i, \text{SES} = j)\} \forall i, j \in \{0, 1\}$$

• Optimization problem:

 $\underset{\theta_{\text{College}}}{\text{argmin}} \left[P(\text{Job}|\text{Talent} = 0, \text{SES} = 0) - P(\text{Job}|\text{Talent} = 0, \text{SES} = 1) \right]^2 +$

 $[P(Job|Talent = 1, SES = 0) - P(Job|Talent = 1, SES = 1)]^{2}$

s.t. $\theta_{i,j} \in \{0,1\} \forall i,j \text{ and } f_{\text{feasibility}}(\theta_{\text{College}}) \leq c_{\text{feasibility}}$



• Admissions policy is θ_{College}

$$\theta_{\text{College}} = \{\theta_{ij} = P(\text{College}|\text{Test} = i, \text{SES} = j)\} \forall i, j \in \{0, 1\}$$

• Optimization problem:

 $\underset{\theta_{\text{College}}}{\operatorname{argmin}} \left[P(\text{Job}|\text{Talent} = 0, \text{SES} = 0) - P(\text{Job}|\text{Talent} = 0, \text{SES} = 1) \right]^2 +$

 $[P(Job|Talent = 1, SES = 0) - P(Job|Talent = 1, SES = 1)]^2$

s.t. $\theta_{i,j} \in \{0,1\} \forall i,j \text{ and } f_{\text{feasibility}}(\theta_{\text{College}}) \leq c_{\text{feasibility}}(\theta_{\text{College}})$



Before and After RAWLSNET



Probability of desired outcome **before** applying RAWLSNET



Probability of desired outcome after applying RAWLSNET

RAWLSNET with feasibility constraints



Probability of desired outcome **after** applying RAWLSNET with **feasibility constraints**

Uses of RAWLSNET

- 1. It can be used to generate "aspirational" data: synthetic data from ideally fair circumstances.
- 2. It can aid policy makers in decision-making.
 - Example: RAWLSNET might be used to advise acceptance decisions for a college admissions committee.
 - Assuming distribution of talent is available in one of these ways, RAWLSNET will calculate the acceptance rates for applicants from different groups needed to satisfy FEO.

A Proposal for Decreasing Geographical Inequality in College Admissions

- Talent is everywhere
- Can we use zip codes and merit to enhance diversity?
- We proposed an algorithm in 2014
 - <u>http://fitelson.org/tie.pdf</u>
- In 2021, it was adapted by Boston Latin School for admissions decisions

for z = 1 to n do Iterate over all the zip+4 codes if $(w_z = 0)$ and $(R_{vs}^2 > 0)$ then $R_{vx}^{r} > 0$ presently but not previously else if $(w_2 > 0)$ and $(R'_{22} = 0)$ then $R'_{va} > 0$ previously but not presently. $W_{*} := ((1 - c) \times W_{*})$ 10: 11: else if $(w_z > 0)$ and $(R'_{vy} > 0)$ then $R_{vs}^{2} > 0$ previously and presently $W_{2} := ((1 - c) \times W_{2}) + 1$ end if end for 15: end for 16: for z = 1 to n do Normalize weights such that $w_z \in [0, 1]$. 17: W. := max((w,) 18: end for

1: c := 10-6

3: W. = 0

4: end for

2: for z = 1 to n do

for y = 1 to m do

Algorithm 1 Calculating weights w_r for the weighted average of the $\{R_{ire}^{t}\}$

Set exponentially decaying constant.

Consider data from the past m years.

Initialize popularity weight w, of zip+4 code z.



Danielle Allen



Branden Fitelson

T. Eliassi-Rad, B. Fitelson. A Proposal for Decreasing Geographical Inequality in College Admissions. Appendix of Chapter 12 (Talent is Everywhere by Danielle Allen) in *The Future of Affirmative Action*, Eds: J. Renker and J. Miller, The Century Foundation Press, 2014.



When data are fair, but the model isn't

<u>Question</u>: Given a fair data distribution and the structure of a BN, does maximum likelihood estimation learn parameters that produce a fair posterior distribution (i.e., one which preserves the fairness in the data distribution)?

When data are fair, but the model isn't

<u>Question</u>: Given a fair data distribution and the structure of a BN, does maximum likelihood estimation learn parameters that produce a fair posterior distribution (i.e., one which preserves the fairness in the data distribution)?

<u>Answer:</u> Not necessarily. It depends on (1) the correctness of the BN's structure, (2) the faithfulness of the learned joint distribution to the BN's structure, (3) the correctness of the learned joint distribution, and (4) whether one is only interested in asymptotic behavior.

Model Cards for Model Reporting

by Margaret Mitchell et al.

https://arxiv.org/abs/1810.03993

Model Card - Smiling Detection in Images

Model Details

- · Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use

- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics

- Evaluation metrics include False Positive Rate and False Negative Rate to measure disproportionate model performance errors across subgroups. False Discovery Rate and False Omission Rate, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).
- Training Data Evaluation Data
 CelebA [36], training data split.
 CelebA [36], test
 - CelebA [36], test data split.
 - Chosen as a basic proof-of-concept.
- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations

Ethical Considerations

- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.



0

0.00 0.02 0.04 0.06 0.08 0.10 0.12 0.14

young

female

male

Model Cards for Model Reporting

by Margaret Mitchell et al.

https://arxiv.org/abs/1810.03993

Model Card - Smiling Detection in Images

Model Details

- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

ntended Use

images; augmentative applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.

- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors

- Description shown problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and woung/old age. Further details available in [36].

Metrics[®]

- Evaluation metrics include False Positive Rate and False Negative Rate to measure disproportionate model performance errors across subgroups. False Discovery Rate and False Omission Rate, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of "fairness" in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR,
- CelebA [36], training data split.
 CelebA [36], training data split.
 - Chosen as a basic proof-of-concept.
- Ethical Considerations tebrities). No new information
- Caveats and Recommendations

Caveats and Recommendations^{of disproportionate errors [5]}.

spectrum of genders.

Ethical Consideration

 An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Quantitative Analyses







	False Omission Rate @ 0.5
old-male	0
old-female	0
ung-female	-0-
oung-male	-0
blo	0
young) O I
male	0
female	0
all	0
0.0	00 0.02 0.04 0.06 0.08 0.10 0.12 0.14

vo

life2vec: vector representations of human lives



Germans Savcisens, et al. Using sequences of life-events to predict human lives. *Nature Computational Science*, 2023. <u>https://doi.org/10.1038/s43588-023-00573-5</u>









How do AI researchers describe and respond to the negative impact of their work on society?

D. Liu, P. Nanayakkara et al. Examining Responsibility and Deliberation in AI Impact Statements and Ethics Reviews. In *Proceedings of the 2022 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society* (AIES), August 2022.

Examining Responsibility and Deliberation in Al Impact Statements and Ethics Reviews in NeurIPS 2021

Impact Statements	Ethics Reviews
Lack of agency	• Issues
Adversarial users	Policy vs. non-policy
 Improper input or application 	• Scope
Denying responsibility	
Explicitly by minimizing negative	Recommendations
societal impact	 Identification vs. mitigation
 Implicitly by not proposing 	
mitigation strategies	Interaction: Justification mechanisms
Assigning responsibility	Citing
To practitioners	Retracting
To the subfield	
To future work	



Nanayakkara

How do AI researchers describe and respond to the negative impact of their work on society? Badly.

D. Liu, P. Nanayakkara et al. <u>Examining Responsibility and Deliberation in AI Impact Statements and Ethics Reviews</u>. In *Proceedings of the 2022 AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society* (AIES), August 2022.

Machine learning life-cycle





C. Wagner, M. Strohmaier, A. Olteanu, E. Kiciman, N. Contractor, T. Eliassi-Rad. Measuring Algorithmically Infused Societies. Nature, 595: 197-204, 2021. https://doi.org/10.1038/s41586-021-03666-1



Claudia Wagner

Markus Strohmaier





Alexandra Olteanu

Emre Kiciman



Noshir Contractor

Three challenges

• Insufficient quality of measurements

• Complex consequences of (mis)measurements

• Limits of existing social theories

How to address these challenges?

- Insufficient quality of measurements
 - Triangulate data to examine the measurement quality
 - Develop guidelines and best practices
- Complex consequences of (mis)measurements
 - Reflect on what to measure and what not to measure
 - Develop professional norms.
- Limits of existing social theories
 - Integrate data and measurements into the theory construction process
 - Establish transparent, participatory processes for examining algorithmically infused societies

ML Life-cycle



Image by Jennifer Wortman Vaughan

Complex System



Image by E Bozzarelli

ML Life-cycle



Image by Jennifer Wortman Vaughan





JAMES LADYMAN & KAROLINE WIESNER

what.

Image by E Bozzarelli



Image by E Bozzarelli

Representations of Complex Systems





Leo Torres



Ann Sizemore Blevins



Danielle Bassett

L. Torres, A. Sizemore Blevins, D. S. Bassett, T. Eliassi-Rad: **The Why, How, and When of Representations for Complex Systems**. *SIAM Review* 63(3): 435-485 (2021). <u>https://doi.org/10.1137/20M1355896</u>



- There is no perfect way to analyze a complex system.
- Modeling decisions made when examining a data set from one system are not **necessarily** ...
 - transferable to another system,
 - or even to another data set from the same system.
Prototypical analysis pipeline for complex systems



Prototypical analysis pipeline for complex systems



• Subset: Are subsets of sets implied?



• Subset: Are subsets of sets implied?



• Temporal: Are walks Markovian?



• Subset: Are subsets of sets implied?



• Temporal: Are walks Markovian?



• Spatial: Are nearby nodes likely to connect?



Are subsets of related nodes necessarily related?



Are subsets of related nodes necessarily related?



Will incorporating temporal dependencies yield more accurate representations?



Spatial dependencies within a system can complicate our representations of the data



External sources of dependencies

- Data availability
 - However big, your data is incomplete.
 - KDD'19 tutorial on network discovery: <u>https://eliassi.org/kdd19tut.html</u>
- Data acquisition and processing
 - Transitivity dependency in projected bipartite graphs does not show up in unipartite graphs.
 - [Gupte & Eliassi-Rad, WebSci'12]
- Research question
 - The question should influence the definition of a relation.
 - Could a common food have caused a disease outbreak?

Understanding system dependencies is a first step in the complex system analysis pipeline



Three types of frameworks composed from nodes and relations



Co-authorship data: Each framework offers its own perspective

Data	QUESTION	Framework	REPRESENTATION
Paper : Authors $p_1 : a_1, a_2$ $p_2 : a_2, a_4$ $p_2 : a_1, a_2, a_3$	Has this pair of authors written together?	\rightarrow Graph \longrightarrow	
$p_4: a_3, a_4$ Authors $a_1 a_2 a_3 a_4$ ρ_1	Has this set of authors	$\rightarrow Simplicial \longrightarrow Complex$	a_1 a_2 a_3 a_4 a_3 a_4 a_3 a_4 a_3 a_4 a_3 a_4 a_3 a_4 a_3 a_4 a_3 a_4 a_4 a_4 a_5 a_4 a_5 a_4 a_5 a_4 a_5 a_4 a_5 a_4 a_5 a_5 a_4 a_5
	Has this set of authors	\rightarrow Hypergraph \longrightarrow	a_1 a_2 a_3 a_4 a_3 a_4 a_3 a_3 a_4

Tenure-track positions are advantageous positions in society

- Getting a tenure-track position depends on many factors
 - School/department: ranking and reputation
 - Advisor(s): professional standing
 - Publications: where and how many
 - Letter writers: professional standing
 - ...
- Letter writers (at least in STEM fields) are often co-authors
- We study the 'who you know' effect
- Assumption: Researchers who collaborate with prominent faculty may receive unfair advantages when applying for faculty positions



'Who you know' effect



- Samantha Dies
- We study CS faculty placement using graph and hypergraph representations of the temporal co-authorship networks.
- We use GNNs to capture the signals in these networks and predict whether a researcher is hired at a high, medium, or low ranked university based on their co-author network leading up to their hire.
- Preliminary results:
 - There is more signal in the exclusivity that hypergraphs capture than pairwise interactions of a simple graph.
 - Publishing within a tightly-knit community of prominent and productive researchers result in a hiring advantage.

Interventions

- So what? Tell me something I don't know.
- What are the number and types of additional collaborations which will improve a hiring candidate's likelihood of being hired at a high-ranked university?



Samantha Dies



Choosing a framework marks the second step of our analysis pipeline











The three frameworks and their analyses can offer different perspectives on a complex system



Core-periphery

Globally circular

Two communities

Variations

- Directed
- Weighted
- Dynamic
- Multiplex
- Multilayer
- Higher-order



Image from Kinsley AC, Rossi G, Silk MJ and VanderWaal K (2020) Multilayer and Multiplex Networks: An Introduction to Their Use in Veterinary Epidemiology. *Frontiers in Veterinary Science* 7:596. doi: 10.3389/fvets.2020.00596

Some of our efforts

- Talent is everywhere
 - Allen, Eliassi-Rad, Fitelson. The Future of Affirmative Action, 2014
 - Liu et al. Al, Ethics, and Society 2021
- Democratic backsliding
 - Wiesner et al. European J. of Physics 2019
 - Eliassi-Rad et al. Humanities & Social Sciences Communications 2020
- Information access equality
 - Wang, Varol, Eliassi-Rad. Applied Network Science 2022
- COVID-19 & the US criminal justice systems
 - Klein et al. *Nature* 2023
- Interventions on academic hiring
 - Dies et al. Conference on Complex Systems 2023
- Using sequences of life-events to predict human lives
 - Germans Savcisens, et al. Nature Computational Science 2023

Some of our efforts

- Talent is everywhere
 - Allen, Eliassi-Rad, Fitelson. The Future of Affirmative Action, 2014
 - Liu et al. Al, Ethics, and Society 2021
- Democratic backsliding
 - Wiesner et al. European J. of Physics 2019
 - Eliassi-Rad et al. Humanities & Social Sciences Communications 2020
- Information access equality
 - Wang, Varol, Eliassi-Rad. *Applied Network Science* 2022
- COVID-19 & the US criminal justice systems
 - Klein et al. Nature 2023
- Interventions on academic hiring
 - Dies et al. Conference on Complex Systems 2023
- Using sequences of life-events to predict human lives
 - Germans Savcisens, et al. Nature Computational Science 2023

Some of our efforts

- Talent is everywhere
 - Allen, Eliassi-Rad, Fitelson. The Future of Affirmative Action, 2014
 - Liu et al. Al, Ethics, and Society 2021
- Democratic backsliding
 - Wiesner et al. European J. of Physics 2019
 - Eliassi-Rad et al. Humanities & Social Sciences Communications 2020
- Information access equality
 - Wang, Varol, Eliassi-Rad. Applied Network Science 2022
- COVID-19 & the US criminal justice systems
 - Klein et al. *Nature* 2023
- Interventions on academic hiring
 - Dies et al. Conference on Complex Systems 2023
- Using sequences of life-events to predict human lives
 - Germans Savcisens, et al. Nature Computational Science 2023

Democracy in the balance

- According to The Economist (13 November 2023), 76 countries will hold elections in 2024.
- That is equivalent to more than half of the world's population going to the polls in 2024.
- EIU has a Democracy Index for 71 of the 76 countries.
 - Of these, 43 countries (more than 60%) have free and fair elections.
 - In the other 28 countries, this is not the case.



September 2020



Democracy has all the features of a complex system.

Image by E Bozzarelli



- Randomness of interactions is important to self-organization
- Democracy requires an unstructured exchange of opinions and ideas between citizens



- Randomness of interactions is important to self-organization
- Democracy requires an unstructured exchange of opinions and ideas between citizens
- <u>Caveat</u>: Instability arises when randomness increases beyond a critical level
- Instability: chaos and collapse in terms of consensual norms



- Instability: chaos and collapse in terms of consensual norms
- "... propels people into accommodating "lying demagogues" because their brazen lies signal opposition to the disdained establishment." Hahl et al. (2018)



Democratic backsliding

Evolution of democracy by category, 2008-21

(Index score out of 10, 10 being best)

Functioning of government Electoral process and pluralism Political participation - Political culture Civil liberties 6.5 6.0 5.5 5.0 4.5 4.0 -2008 10 11 12 13 15 16 17 18 19 20 21 14 Source: EIU.

Democratic backsliding is, in part, the result of instability in the democratic system.

Al technology that amplifies misinformation and disinformation increases instability in democracy.



(Temporal) Stability vs. (Modal) Robustness

Stability

- Persistence over time
- Resistance to perturbation or change
- Example: a rock on top of a peak is <u>not</u> stable



Robustness

- Insensitivity or independence of the behavior of the system to changes in possible microscopic realization
- Example: *k*NN is <u>not</u> robust


(Temporal) Stability vs. (Modal) Robustness

Stability

- Persistence over time
- Resistance to perturbation or change
- Example: a rock on top of a peak is <u>not</u> stable



Robustness

- Insensitivity or independence of the behavior of the system to changes in possible microscopic realization
- Example: *k*NN is <u>not</u> robust



A stable system need not be robust nor a robust system stable.

Understanding democracy as a complex system enables us to make policy recommendations to counteract democratic backsliding (including ones caused by AI technology).

Policy recommendations



- 1. Entrench diversity by regulation
- 2. Monitor feedback
- 3. Ensure connectivity



- 1. Recruit credible communicators in estranged regions
- 2. Recognize limits to message control
- 3. Emphasize persistence and limits to forecasting

Disclaimer on policy recommendations

- Our recommendations do not form a hierarchy
- Their effectiveness is context dependent
 - The way in which these recommendations can be put into practice will differ from country to country
- Our recommendations are also not independent of each other
 - Example: failure in recruiting credible communicators increases the likelihood of failure in recognizing limits of message control
- Democracy is an evolving project



OPEN ACCESS IOP Publishing

European Journal of Physics

Eur. J. Phys. 40 (2019) 014002 (17pp)

https://doi.org/10.1088/1361-6404/aaeb4d

Stability of democracies: a complex systems perspective

K Wiesner^{1,2}, A Birdi³, T Eliassi-Rad⁴, H Farrell⁵, D Garcia^{2,6}, S Lewandowsky⁷, P Palacios⁸, D Ross^{9,10,11}, D Sornette¹² and K Thébault¹³

¹School of Mathematics, University of Bristol, United Kingdom ²Complexity Science Hub Vienna, Austria ³ Department of Economics, University of Bristol, United Kingdom ⁴Network Science Institute, College of Computer and Information Science, Northeastern University, Boston, United States of America ⁵Department of Political Science, George Washington University, Washington, D.C., United States of America ⁶Section for Science of Complex Systems, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Austria ⁷School of Psychological Science, University of Bristol, United Kingdom ⁸ Department of Philosophy, University of Salzburg, Austria ⁹School of Sociology, Philosophy, Criminology, Government, and Politics, University College Cork, Ireland ¹⁰ School of Economics, University of Cape Town, South Africa ¹¹Center for Economic Analysis of Risk, Georgia State University, United States of America

¹² Department of Management, Technology and Economics, ETH Zurich, Switzerland ¹³ Department of Philosophy, University of Bristol, United Kingdom

E-mail: k.wiesner@bristol.ac.uk

Received 25 June 2018, revised 21 September 2018 Accepted for publication 25 October 2018 Published 27 November 2018



Abstract

The idea that democracy is under threat, after being largely dormant for at least 40 years, is looming increasingly large in public discourse. Complex systems theory offers a range of powerful new tools to analyse the stability of social institutions in general, and democracy in particular. What makes a democracy stable? And which processes potentially lead to instability of a democratic system? This paper offers a complex systems perspective on this question, informed by areas of the mathematical, natural, and social sciences. We

COMMENT

Check for updates

https://doi.org/10.1057/s41599-020-0518-0 OPE

What science can do for democracy: a complexity science approach

Tina Eliassi-Rad¹, Henry Farrell², David Garcia^{3,4}, Stephan Lewandowsky ⁵, Patricia Palacios⁶, Don Ross^{7,8,9}, Didier Sornette¹⁰, Karim Thébault¹¹ & Karoline Wiesner ¹²⁸³

Political scientists have conventionally assumed that achieving democracy is a one-way ratchet. Only very recently has the question of "democratic backsliding" attracted any research attention. We argue that democratic instability is best understood with tools from complexity science. The explanatory power of complexity science arises from several features of complex systems. Their relevance in the context of democracy is discussed. Several policy recommendations are offered to help (re)stabilize current systems of representative democracy.

¹Network Science Institute, College of Computer and Information Science, Northeastern University, Boston, MA, USA ²Department of Political Science, George Washington University, Washington, DC, USA ³Section for Science of Complex Systems, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria.⁴ Complexity Science Hub Vienna, Vienna, Austria.⁵School of Psychological Science, University of Bristol, DK.⁶Department of Philosophy, University of Salzburg, Salzburg, Austria.⁷School of Science, University of Politics, University of Interna, Vienna, Austria, ⁵School of Science, University of Salzburg, Salzburg, Cape Town, Cape Town, South Africa.⁹Center for Economic Analysis of Rsk, Georgia State University, Atlanta, CA, USA.¹⁰Department of Management, Technology and Economics, ETH Zurich, Switzerland.¹¹Department of Philosophy, University of Strito, Bristol, UK.¹⁰School of Homentics, University of Bristol, UK.⁸Benait, Kweisengibhristolacuk

-DMANITHY AND LOCIAL SCIENCES COMMUNICATIONS (2020)7:30 [https://doi.org/10.1057/s41599-020-0518-0

Downloadable from http://eliassi.org/pubs.html



Machine learning (ML) systems are not islands.

They are part of broader complex systems.

To understand and mitigate the risks and harms of using ML, we must remove our optimization blinders & study the broader complex systems in which ML systems operate.

Thank you!

Any questions?

Some of our efforts

- Talent is everywhere
 - Allen, Eliassi-Rad, Fitelson. The Future of Affirmative Action, 2014
 - Liu et al. Al, Ethics, and Society 2021
- Democratic backsliding
 - Wiesner et al. European J. of Physics 2019
 - Eliassi-Rad et al. Humanities & Social Sciences Communications 2020
- Information access equality
 - Wang, Varol, Eliassi-Rad. Applied Network Science 2022
- COVID-19 & the US criminal justice systems
 - Klein et al. Nature 2023
- Interventions on academic hiring
 - Dies et al. Conference on Complex Systems 2023
- Using sequences of life-events to predict human lives
 - Germans Savcisens, et al. Nature Computational Science 2023

Complex networks are ubiquitous





Biological networks

Information Networks



Food Web



Contagion of TB



Social and information networks







Common mechanisms and properties limiting minority nodes' access to information



L. Espín-Noboa, C. Wagner, M. Strohmaier, F. Karimi. Inequality and Inequity in Network-based Ranking and Recommendation Algorithms. *Scientific Reports* 12 (1): 1-14 (2022)





Xindi Wang Onur Varol

What is the interplay between **network structure** and the spreading process for information access equality?

X. Wang, Onur Varol, T. Eliassi-Rad. Information Access Equality on Generative Models of Complex Networks. Applied Network Science, Volume 7, Article 54, 2022. https://doi.org/10.1007/s41109-022-00494-8

Growth mechanisms

- Majority/minority dichotomy
- Homophily
- Preferential attachment
- Diversity

Growth mechanisms

- Majority/minority dichotomy
- Homophily
- Preferential attachment
- Diversity

Spreading processes

- Simple vs. complex contagions
- Symmetric vs. asymmetric transmission rates
- Various seeding conditions

• Spreading processes can take different times in different complex networks

- Spreading processes can take different times in different complex networks
- Measure information access equality at different stages of the spreading process

- Spreading processes can take different times in different complex networks
- Measure information access equality at different stages of the spreading process
- Normalize fraction of nodes in state I (Infected) at each time step t by the length of the spreading process T to obtain I(t/T)

$$\Delta I(t/T) = \frac{I_{maj}(t/T) - I_{min}(t/T)}{I_{maj}(t/T) + I_{min}(t/T)} \in [-1, 1]$$

- Spreading processes can take different times in different complex networks
- Measure information access equality at different stages of the spreading process
- Normalize fraction of nodes in state I (Infected) at each time step t by the length of the spreading process T to obtain I(t/T)

$$\Delta I(t/T) = \frac{I_{maj}(t/T) - I_{min}(t/T)}{I_{maj}(t/T) + I_{min}(t/T)} \in [-1, 1]$$

$\Delta I(t/T)$	Meaning
< 0	Minority group is at an advantage
= 0	Equality
> 0	Majority group is at an advantage

1. Information access equality is a complex interplay between network structures and the spreading processes.

- 1. Information access equality is a complex interplay between network structures and the spreading processes.
- 2. There is a trade-off between equality and efficiency of information access under certain circumstances (e.g., low inter-group edges and asymmetric transmission).

- 1. Information access equality is a complex interplay between network structures and the spreading processes.
- 2. There is a trade-off between equality and efficiency of information access under certain circumstances (e.g., low inter-group edges and asymmetric transmission).
- Spreading process features are statistically significant (p-value ≤ 0.05) when it comes to information access equality.

- 1. Information access equality is a complex interplay between network structures and the spreading processes.
- 2. There is a trade-off between equality and efficiency of information access under certain circumstances (e.g., low inter-group edges and asymmetric transmission).
- 3. Spreading process features are statistically significant (p-value ≤ 0.05) when it comes to information access equality.
- 4. Network features are not always statistically significant. But two network features stand out w.r.t. information access equality: degree inequality and network distance.

Downstream impact

- Our findings can be used to recommend connections that steer an online social network toward information access equality for a given topic.
- Ideally, such recommendation systems will first classify the spreading process, then based on the classification recommend new connections to their users.





Some of our efforts

- Talent is everywhere
 - Allen, Eliassi-Rad, Fitelson. The Future of Affirmative Action, 2014
 - Liu et al. Al, Ethics, and Society 2021
- Democratic backsliding
 - Wiesner et al. European J. of Physics 2019
 - Eliassi-Rad et al. Humanities & Social Sciences Communications 2020
- Information access equality
 - Wang, Varol, Eliassi-Rad. *Applied Network Science* 2022
- COVID-19 & the US criminal justice systems
 - Klein et al. Nature 2023
- Interventions on academic hiring
 - Dies et al. Conference on Complex Systems 2023
- Using sequences of life-events to predict human lives
 - Germans Savcisens, et al. Nature Computational Science 2023

Prison population fell during COVID-19

- Almost all courts shut down, reducing the admission rate by 70 percent.
- 2. Prisoners were released in response to the pandemic.





B. Klein et al. <u>The COVID-19 Pandemic Amplified long-standing Racial Disparities in the United States Criminal</u> <u>Justice System</u>. *Nature*, 2023.

Prison population fell during COVID-19

- The fraction of incarcerated Black and Latino people increased.
- But in most states, the released prisoners were not disproportionately white.





Dynamics of the US prison population

- 1. Non-white people tend to get longer sentences than white people for the same crimes.
 - With fewer prisoners coming in over the course of the pandemic and with white prisoners disproportionately serving shorter sentences, the population skewed Black and Latino.
- 2. During the pandemic, prosecutors pushed hard for pre-trial plea deals to complete cases.
 - Plea deals result in a disproportionate number of Black defendants spending time in prison.
- 3. Decreasing the flow of new admissions increased the non-white population.
 - The Black-white disparity in new prison admissions is typically a ratio of 2:1, whereas it is closer to 6:1 for the total incarcerated.

Why should you care?

- There is still a large backlog of cases in the criminal legal system stemming from delays in the early stages of the pandemic
- The structural problem in the U.S. criminal legal system will continue to worsen unless we address the sentencing inequities in the legal system and work toward reforms that will produce a more equitable and just system