nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00793-y

Mitigating allocative tradeoffs and harms in an environmental justice data tool

Received: 25 May 2023

Accepted: 9 January 2024

Published online: 16 February 2024

Check for updates

Benjamin Q. Huynh[1,8], Elizabeth T. Chin[2,8], Allison Koenecke[3], Derek Ouyang[4], Daniel E. Ho[4,5,6], Mathew V. Kiang[7,9] & David H. Rehkopf[7,9]

Neighbourhood-level screening algorithms are increasingly being deployed to inform policy decisions. However, their potential for harm remains unclear: algorithmic decision-making has broadly fallen under scrutiny for disproportionate harm to marginalized groups, yet opaque methodology and proprietary data limit the generalizability of algorithmic audits. Here we leverage publicly available data to fully reproduce and audit a large-scale algorithm known as CalEnviroScreen, designed to promote environmental justice and guide public funding by identifying disadvantaged neighbourhoods. We observe the model to be both highly sensitive to subjective model specifications and financially consequential, estimating the effect of its positive designations as a 104% (62–145%) increase in funding, equivalent to US$2.08 billion (US$1.56–2.41 billion) over four years. We further observe allocative tradeoffs and susceptibility to manipulation, raising ethical concerns. We recommend incorporating technical strategies to mitigate allocative harm and accountability mechanisms to prevent misuse.

As decision-making algorithms continue to be adopted for a variety of high-impact applications, many of them have been found to disproportionately harm marginalized populations, as evidenced by algorithmic audits[1–6]. In particular, area-based measures to identify disadvantaged neighbourhoods have recently become widespread for tasks such as allocating vaccines, assessing social vulnerability, and healthcare cost adjustment, intended to optimize an equitable distribution of resources[7–10]. However, their potential for allocative harm—the withholding of resources from specific subpopulations[11]—is not well understood, and it remains unclear how different subpopulations may be disproportionately impacted by the design of such area-based models.

The California Community Environmental Health Screening Tool (CalEnviroScreen) is a data tool that designates neighbourhoods as eligible for capital projects and social services funding, and is intended to promote environmental justice. CalEnviroScreen's model output is used to designate 'disadvantaged communities', for which 25% of proceeds from California's cap-and-trade programme are earmarked. CalEnviroScreen also directly influences funding from a variety of public and private sources, and is reported to have directed an estimated US$12.7 billion in funding[12]. The funding targets of the tool are varied, including programmes for affordable housing, land-use strategies, agricultural subsidies, wildfire risk reduction, public transit and renewable energy. Similar data tools are in use or development at the federal and state levels across the United States[13].

CalEnviroScreen ranks each census tract in the state according to its level of marginalization in terms of environmental conditions and population characteristics. The algorithm does so by aggregating publicly available tract-level data into a single score, based on variables

[1]Department of Environmental Health and Engineering, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [3]Department of Information Science, Cornell University, Ithaca, NY, USA. [4]RegLab, Stanford University, Stanford, CA, USA. [5]Stanford Law School, Stanford, CA, USA. [6]Department of Political Science, Stanford University, Stanford, CA, USA. [7]Department of Epidemiology & Population Health, Stanford University School of Medicine, Stanford, CA, USA. [8]These authors contributed equally: Benjamin Q. Huynh, Elizabeth T. Chin. [9]These authors jointly supervised this work: Mathew V.Kiang, David H. Rehkopf. e-mail: bhuynh@jhu.edu

https://doi.org/10.1038/s42256-024-00793-y

from four categories: environmental exposures, environmental effects, sensitive populations and socioeconomic factors. Tracts in the top 25% of scores are designated as disadvantaged communities, representing ~10 million residents for whom earmarked funding is made available.

Screening tools like CalEnviroScreen have been criticized for being in tension with the principles of environmental justice, defined broadly as a movement to address systemic environmental harms faced by marginalized populations through mechanisms such as equitable resource allocation and inclusive decision-making[14]. On the one hand, such tools may distributively advance environmental justice by allocating resources to marginalized communities, but on the other, critics contend that subjective, state-run decision-making lacks accountability for affected communities, particularly as the state itself bears responsibility for perpetuating environmental injustices[14–16]. Consequently, such screening tools may present their algorithmic output as objective truth and place marginalized communities in competition for limited funds[16,17]. Algorithmic audits are therefore necessary to identify the extent to which the design of tools like CalEnviroScreen can impact different communities.

Audits of large-scale algorithms are often limited to observing 'black box' outputs, as individual-level data privacy requirements and proprietary pipelines prevent comprehensive audits of algorithmic systems[1,2,18]. By contrast, we are able to fully reproduce and test changes to CalEnviroScreen's model due to its population-level usage of publicly available data, potentially enabling generalization of our findings to similar large-scale algorithms. In this Article we investigate the inner workings of CalEnviroScreen, characterizing model sensitivity, funding impact, ethical concerns and avenues for harm-reduction (Extended Data Fig. 1).

Model sensitivity and funding impact

The CalEnviroScreen model is highly sensitive to change. We found that 16.1% of all tracts could change designation based on small alterations to the model. This represents high designation variation given that only 25% of all tracts receive designation (Fig. 1 and Extended Data Table 1). These large fluctuations in designation are solely due to varying subjective model specifications such as health metrics, pre-processing methods and aggregation methods[19]. For example, changing pre-processing methods—switching from a percentile ranking to a more commonly used method like *z*-score standardization—led to a 5.3% change in designated tracts.

In the absence of a ground-truth variable or validation metric (that is, a concrete ability to quantify the true value of environmental harm in California), model sensitivity represents the ambiguity across alternative specifications, enabling an uncertainty assessment of the model outputs[19–22]. For example, we observe high levels of model sensitivity at the designation threshold (75th percentile), where the predicted tract ranking could vary across models by 44 percentile ranks (Fig. 1). Even tracts ranked as low as the bottom 5th percentile could be eligible under slightly different models. We observe lower, yet still substantial, model sensitivity at the 95th percentile, where the predicted range is 18 percentile ranks. Given this variability in ranking certainty, dichotomizing designation may present a false sense of precision, leading to funding decisions based on unstable information.

Receiving algorithmic designation is financially consequential. We estimated through a causal analysis that the effect of receiving designation from the algorithm is a 104% (95% confidence interval, 62–145%) increase in funding, equivalent to US$2.08 billion (US$1.56–2.41 billion) in additional funding over a four-year period for 2,007 tracts (Fig. 2 and Extended Data Table 2). Similarly, among the 400 tracts that would be eligible for designation under an alternative model (described below), we estimated they would have received equivalent to US$632 million ($377–881 million) in additional funding over the same time period.

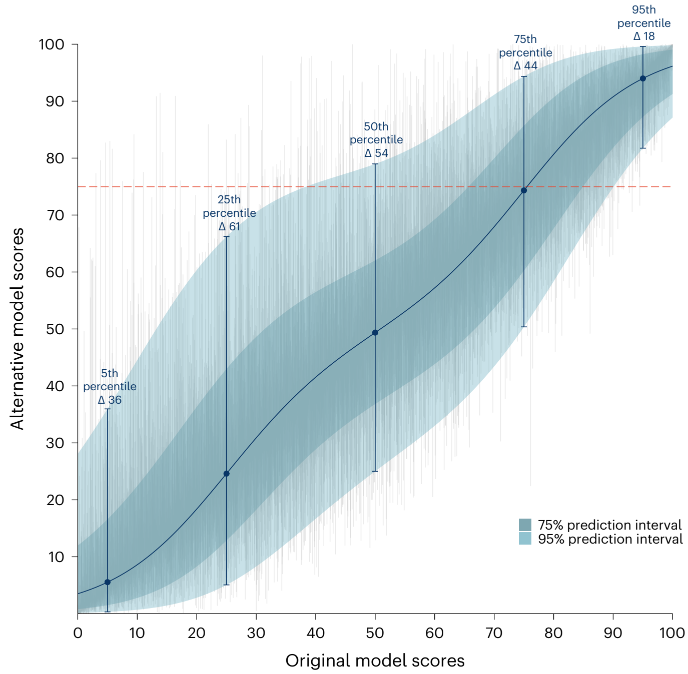


Fig. 1 | CalEnviroScreen's sensitivity to input parameters. The axes denote model scores in terms of percentiles. Grey bars indicate maximum and minimum values from alternative plausible model specifications with varying health metrics, pre-processing methods and aggregation methods. The dashed red line indicates the 75th percentile cutoff score for funding designation. Dots indicate the median predicted amount of model sensitivity at a given percentile, in terms of by how many percentile-ranks a tract can vary. Light shaded portions and error bars indicate 95% prediction intervals. Dark shaded portions indicate 75% prediction intervals (for example, in 95% of predictions, tracts at the 75th percentile can vary their score by 44 percentile-ranks).

Allocative tradeoffs and harms

Under such a model with high uncertainty, every subjective model decision is implicitly a value judgement: any variation of a model could favour one subpopulation or disfavour another. Both the model in its current form and plausible alternative forms can exhibit bias among different subpopulations, illustrating the zero-sum nature of delegating funding allocation to a single model.

To exemplify these challenging tradeoffs, we constructed an alternative model for designation assignment. CalEnviroScreen does not include race in its algorithm, but we are able to assess its impact on race by examining the racial composition of designated tracts. We first changed the pre-processing and aggregation methods to avoid penalizing tracts with extreme levels in variables such as air pollution indicators, and then incorporated a number of additional population health metrics for a more broad definition of vulnerability to environmental exposures. On average, incorporating these changes led to increased designation to tracts with higher levels of racially minoritized people in poverty, but decreased designation among racially minoritized populations overall (Fig. 3 and Extended Data Fig. 2).

In particular, expanding the 'sensitive populations' category of the algorithm presents ethical concerns. The category is represented by three variables: respiratory health, cardiovascular health and low birthweight. It would be sensible to include additional health indicators relevant to environmental exposures, such as chronic kidney disease or cancer[23,24]. The inclusion of such variables, however, would result in the loss of designation for tracts with high Black populations. Because low birthweight disproportionately affects Black infants, the introduction of other variables such as cancer—which also disproportionately affects Black populations albeit to a lesser extent—would reduce the impact of low birthweight on the algorithm's output.

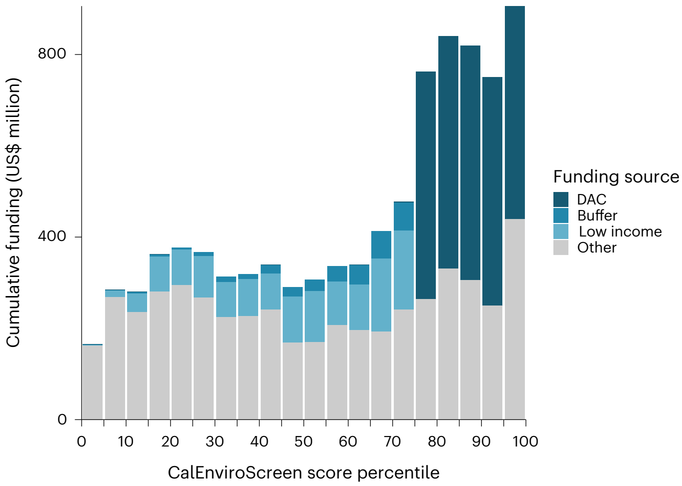


Fig. 2 | Total cumulative funding by California Climate Investments implemented in census tracts by CalEnviroScreen percentile from 2017 to 2021. Dark blue bars indicate funding earmarked specifically for disadvantaged communities (DAC), as determined by the CalEnviroScreen model. Lighter blue bars indicate other earmarked funding (buffer and low income). Grey bars indicate all other funding. Buffer funding is earmarked for low-income communities and households that are not designated as disadvantaged communities, but are within half a mile of a disadvantaged communities census tract. Low-income funding is earmarked for low-income communities and households statewide.

Moreover, we found the existing model to potentially underrepresent foreign-born populations. The model measures respiratory health in terms of emergency-room visits for asthma attacks, which underrepresents groups who use the emergency room less or come from countries where asthma is less prevalent, yet still have other respiratory issues[25,26]. Consequently, using survey data of chronic obstructive pulmonary disease (COPD) to represent respiratory health increases the designation of tracts with foreign-born populations of 30% or higher (Extended Data Figs. 3 and 4).

Critically, the zero-sum nature and high sensitivity of the model are conducive to model manipulability. It is feasible for a politically motivated internal actor, whether subconsciously or intentionally, to prefer model specifications that designate tracts according to a specific demographic, such as political affiliation or race. Through adversarial optimization—optimizing over pre-processing and aggregation methods, health metrics and variable weights—we find that the maximum increase or decrease a model can be manipulated to favour a specific US political party is 39% and 34%, respectively (Fig. 4 and Extended Data Fig. 5). Any efforts to mitigate the harms of allocative algorithms such as CalEnviroScreen thus need to consider both model sensitivity and manipulability.

Mitigation strategies

Because there is no singular 'best' model, we propose assessing robustness via sensitivity analysis and incorporating additional models accordingly. For example, the California Environmental Protection Agency recently decided to honour designations from both the current and previous versions of CalEnviroScreen, effectively taking the union of two different models. This approach reduces model sensitivity by 40.7%, and a three-model approach additionally incorporating designations from our alternative model reduces the model sensitivity by 71.0%. Using multiple models also mitigates allocative harm—by broadening the category of who is considered disadvantaged, different populations are less likely to be in competition with each other for designation.

A potential concern is that increasing the number of designated tracts may dilute earmarked funds for disadvantaged groups. However, incorporating an additional model per our example would only increase the number of tracts by 10%, yet reduce model sensitivity by 51.1%. Doing so would also reduce equity concerns and more accurately represent the uncertainty inherent to designating tracts (consideration should be given as to whether these benefits outweigh the downsides). Furthermore, adding models is only one possible solution. There are many other ways to equitably address decision-making under uncertainty, such as randomizing assignment for tracts near the decision threshold (similar to lottery admissions for educational institutes), aggregating outputs from multiple models into a singular ensemble model, scoring tracts based on both the model output and its respective uncertainty measurement, or funding tracts on a tiered or sliding-scale system weighted by uncertainty measurements instead of using a single hard threshold[27–30].

However, reducing model sensitivity is not a complete solution—transparency and accountability are necessary to reduce harm. The agency developing CalEnviroScreen is active in offering methodological transparency and soliciting feedback, which enables critiques such as ours and promotes public discourse. Agencies developing similar tools to identify disadvantaged neighbourhoods should follow suit. A safeguard like an external advisory committee comprising domain experts and leaders of local community groups could also help reduce harm by identifying ethical concerns that may have been missed internally. It would also promote equitable representation and involvement from the public, aligning with the tool's goal of advancing environmental justice.

Discussion

Our findings are threefold: (1) CalEnviroScreen's model is both sensitive to change and financially consequential; (2) subjective model decisions lead to allocative tradeoffs, and models can be manipulated accordingly; and (3) model sensitivity can be mitigated by accounting for uncertainty in designations, thereby reducing the need for tradeoffs. Concretely, we recommend accounting for uncertainty by incorporating sensitivity analyses and potentially including additional models to increase robustness, and urge for community-based independent oversight.

Our analysis is not a comprehensive audit of CalEnviroScreen. We do not identify every potential flaw or ethical concern of the model, but instead highlight illustrative examples of how model choices can facilitate allocative harm. Only members of a given community can fully know how their respective tracts are represented and affected by the algorithm. We do not advocate for any particular model over another. Such decisions are inherently subjective and should be made in consultation with affected communities and relevant experts. Our estimates of model sensitivity are probably underestimates, as we do not exhaustively specify alternative models. Furthermore, our estimate of the funding impact of algorithmic designation is probably an underestimate, as detailed data on relevant private funding sources are not publicly available. Our estimate of model bias for foreign-born populations may be inaccurate because undocumented immigrants are often underrepresented in census data[31]. We are unable to fully assess model sensitivity to the modifiable areal unit problem[32]—a phenomenon where varying the geographical unit of observation can alter model output—because we only had data to convert a fraction of the variables to a smaller geographical scale (Supplementary Information).

Other limitations of our analysis include the limitations besetting the data tool itself: missingness in data and a lack of random measurement error metrics. For example, indoor air quality or regulatory compliance may be important determinants of environmental risk, but are not included in CalEnviroScreen[33,34]. Similarly, CalEnviroScreen measures outdoor air quality using sensors and interpolation algorithms used to infer air quality for areas between sensors, which may result in noisier estimates for marginalized communities farther from sensors. The degree to which such algorithms can cause allocative harm should

 https://doi.org/10.1038/s42256-024-00793-y

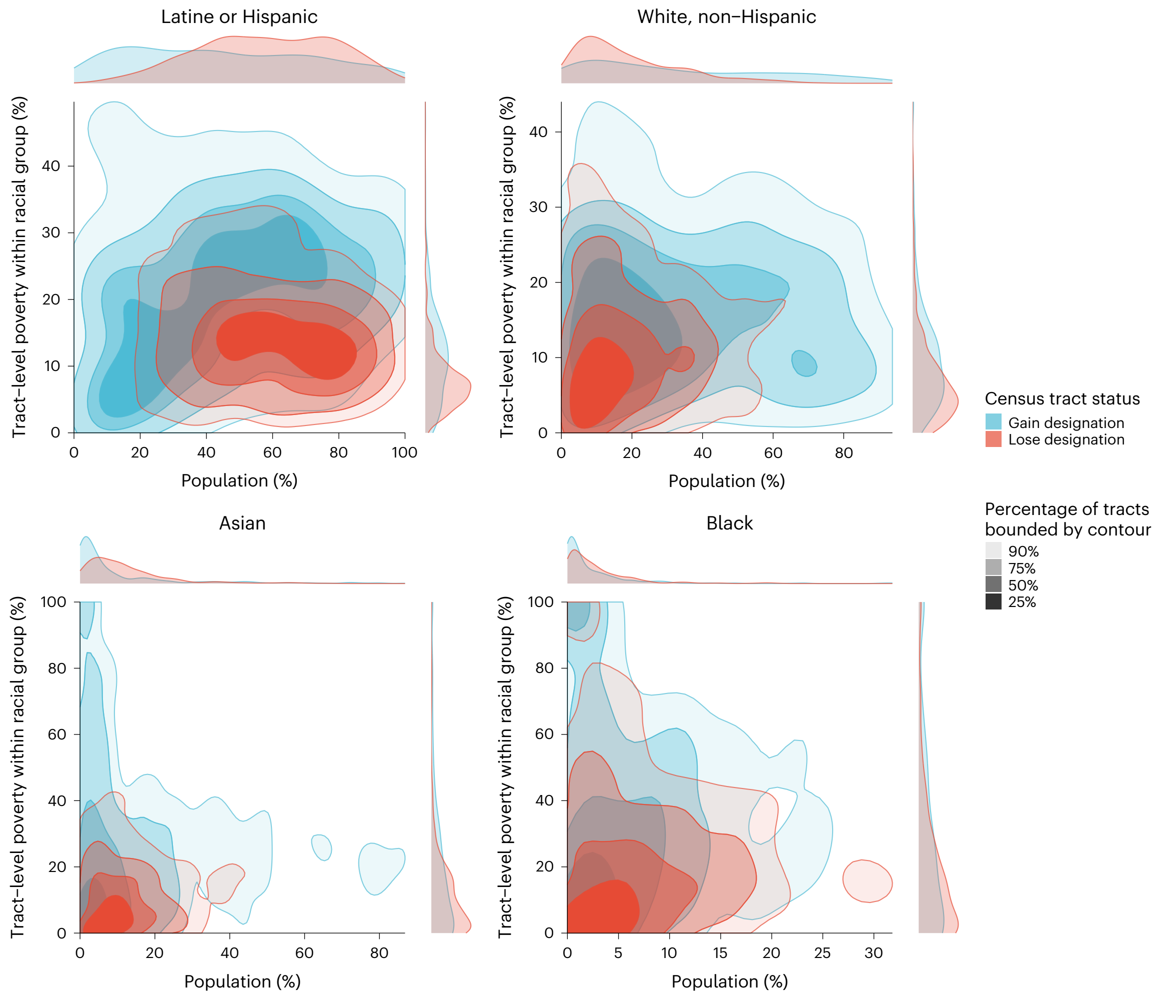


Fig. 3 | Allocative tradeoffs between racially minoritized populations in poverty and racially minoritized populations overall. Comparison of how algorithmically designated tracts are distributed by race and poverty across the current CalEnviroScreen model and an alternative model, among tracts that would change designation status under the alternative model. The alternative model uses a different pre-processing technique, a different aggregation technique, and it incorporates additional population health variables. Red densities indicate tracts that receive designation under the current model but are not designated under the alternative model. Blue densities indicate tracts gaining designation under the alternative model. Contours are calculated as the smallest regions that bound a given proportion of the data.

be examined[35]. Overall, missingness in environmental data is more pronounced in marginalized communities[36]. As CalEnviroScreen's data sources lack random measurement error metrics, our uncertainty estimates only reflect a specific type of uncertainty: model sensitivity, or ambiguity between models[21,22]. Incorporating random measurement error would increase the uncertainty of CalEnviroScreen scores.

Our work draws upon previous literature from a variety of disciplines. Previous studies related to decision science and composite indicator construction recognize model sensitivity, or ambiguity, as a measure of uncertainty, and demonstrate how diversity in modelling assumptions, or worldviews, can improve robustness[19,21,22,37,38]. In the algorithmic fairness literature, uncertainty has been formulated as a driver of algorithmic bias for ranking algorithms, and race has been found to be inferred from models that do not include it as a variable[5,20,39]. Environmental justice theorists have critiqued environmental screening algorithms for extractive data practices, the exclusion of affected communities from decision-making, the subjectivity of outside expertise in allocating community resources, and not including race as a variable[7,16,17,36,40]. Our analysis also draws from frameworks of 'distributive justice', examining how to fairly allocate resources within a society[41–44]. A recent environmental health study has examined how environmental screening tools can be improved to mitigate disparities in air quality[45]. Our analysis contributes to the literature by identifying technical mechanisms by which subjectivity in the model design of environmental screening algorithms contributes to uncertainty in the model output and the potential for allocative harm.

More broadly, our findings illustrate how allocative algorithms can encode unintentional bias into their outputs. Questions of how to allocate scarce resources have always been challenging and subjective, yet delegating allocation to algorithms may erroneously give the appearance of objectivity by obscuring the design choices behind the algorithms[15,17,46]. Any such notion that algorithms are intrinsically objective should be rejected. With increasingly high-dimensional and high-resolution data, unintentional bias will become both more common and more difficult to detect. Both algorithm developers and policymakers should acknowledge the subjective process of algorithm development and work to minimize harm accordingly.

Technical and regulatory solutions will be necessary to address the concerns of allocative harm as algorithms continue to be adopted for policy use. Although the misuse of such tools could exacerbate existing inequities, a careful and community-minded approach can lead to the broad realization of CalEnviroScreen's intended goal—furthering environmental justice and mitigating the harms done to structurally marginalized populations.

https://doi.org/10.1038/s42256-024-00793-y

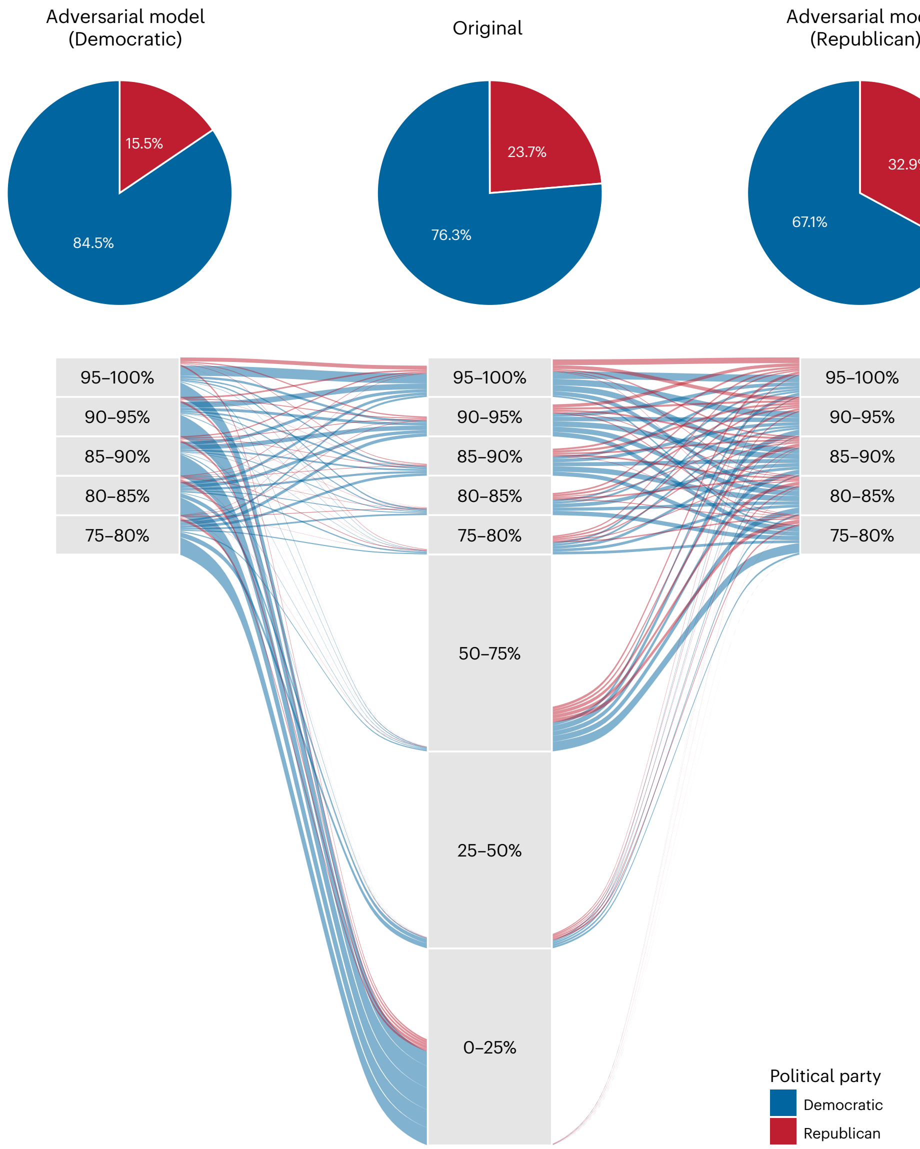


Fig. 4 | Adversarially optimized distribution of algorithmically designated tracts by political party. Columns and flow denote the distribution of tracts designated as disadvantaged by political affiliation, determined by affiliations of district assembly members. The leftmost and rightmost columns are adversarially optimized to increase Democratic and Republican tracts, respectively. The centre column represents the original model. Flows between columns represent changes in binned percentile ranking of census tracts between the original and adversarial models.

Methods

Data

For our sensitivity analyses, we used census tract-level data obtained from the current version of CalEnviroScreen (version 4.0, implemented in late 2021), consisting of 8,035 observations with 21 variables measuring different aspects of environmental exposures and population characteristics. The variables measure ozone levels, fine particulate matter, diesel particulate matter, drinking water contaminants, lead exposure, pesticide use, toxic release from facilities, traffic impacts, cleanup sites, groundwater threats, hazardous waste, impaired waters, solid waste sites, asthma, cardiovascular disease, low birthweight, education, housing burden, linguistic isolation, poverty and unemployment.

For our additional variable analyses, we used the PLACES dataset from the Centers for Disease Control and Prevention to include tract-level variables on estimated prevalence for asthma, cancer, chronic kidney disease and coronary heart disease. We obtained demographic information on tract-level race/ethnicity from the American Community Survey.

For the causal analysis specifically, we examined years 2017–2021 of the California Climate Investments funding dataset, and used CalEnviroScreen 3.0 scores (implemented in 2017), as there are not yet sufficient data for funded projects guided by CalEnviroScreen 4.0. We calculated the amount of funding allocated to each census tract by summing the amount of funding received from different programmes for each tract. For funding projects that were attributed to assembly districts but not specific census tracts, we made conservative assumptions that prioritized non-earmarked funding towards non-priority population tracts. We attributed assembly district-level funding to tracts using the following steps: (1) funds earmarked for 'priority populations' such as disadvantaged tracts were exclusively attributed to their respective tracts within that district; (2) the remaining funds were attributed to non-priority population tracts within the district up to the amount attributed to priority population tracts; (3) any remaining funds after that were distributed equally (more details are provided in the Supplementary Information). To attribute districts to tracts spanning multiple districts, we followed the methodology listed in the

https://doi.org/10.1038/s42256-024-00793-y

California Climate Investments funding dataset—we solely considered them to belong to whichever district contained the largest population. For tracts that were missing relevant block-level population metrics, we assigned them to districts based on whichever district contained more blocks from the given tract. For a single tract that had missing population metrics and the same number of blocks for two districts, we assigned its district based on geographical area.

Algorithmic audit

We first reproduced the original CalEnviroScreen model based on its documentation[47], then validated our reproduction on existing data. We next identified potential issues in the data tool and conceived plausible alternative models. As a general approach, we built alternative models (implementing various small changes to the current CalEnviroScreen model) and evaluated how they differed from the original model to assess the sensitivity of the CalEnviroScreen algorithm to model decisions. Variation was measured in terms of percent change in tracts changing designation, that is, the number of tracts changing designation divided by the total number of tracts multiplied by 100. Details of each step of this approach are given in the following.

We assessed changes to (1) the pre-processing method, (2) the aggregation method and (3) health metrics, all subjective areas for constructing composite indicators. We assessed pre-processing methods by changing the existing pre-processing method—percentile-ranking—to z-score standardization. We assessed the aggregation methods by changing the existing aggregation method—multiplication—to arithmetic mean.

We assessed health metrics based on our concerns of public health biases perpetuated by the algorithm. First, we noted that the existing method of measuring health indicators strictly by emergency-room visits may be skewed towards populations who use the ER disproportionately often, and so we tested including tract-level survey indicators of health in the model, namely asthma and cardiovascular health[26]. Second, we noted that only using asthma as a measure of environmental vulnerability with respect to respiratory health may not be fully reflective of those with respiratory health issues, so we tested including survey indicators for COPD. The inclusion of survey indicators of health were weighted such that categories of respiratory health, cardiovascular health and low birthweight were equally weighted. Finally, we noted that low birthweight, cardiovascular and respiratory issues are not the only health-related ways in which populations may be vulnerable to environmental exposures, and so we tested including other indicators, such as chronic kidney disease and cancer.

Our alternative models were pre-specified and designed based on the changes listed above: changing pre-processing to standardization, changing aggregation to averaging, and including survey indicators of health for cardiovascular health, asthma, COPD, chronic kidney disease and cancer. We evaluated the overall model sensitivity by assessing the different combinations of model specifications by varying pre-processing (z-score standardization versus percentile ranking), aggregation (multiplication versus averaging) and health variables (including versus excluding the additional health variables we specified)—calculating the number of distinct tracts that change designation across all models. We trained a smooth nonparametric additive quantile regression model on the range (that is, minimum and maximum values across models) for each tract to obtain prediction intervals[48].

Empirical strategy

We used a sharp regression discontinuity design with local linear regression as the functional form to estimate the effect of algorithm designation on total funding received[49]. We selected the bandwidth using the Imbens–Kalyanaraman algorithm[50]. The treatment variable was a binary indicator for each tract denoting whether it was designated as disadvantaged by the algorithm. The outcome variable was the log of total funding received per tract. The forcing variable was the CalEnviroScreen percentile rank for each tract. Covariates included the aggregate pollution burden and population characteristics indicators from CalEnviroScreen, and tract-level race and poverty estimates from the American Community Survey. As robustness checks, we estimated the treatment effect with varying bandwidths, functional forms, covariate adjustments and dataset configurations. We also estimated the treatment effect with a propensity score matching approach[51], and a linear model causal forest (Supplementary Information)[52]. All parenthetical values reported in the main text are 95% confidence intervals, and were calculated by multiplying standard errors by the 97.5th percentile point of the standard normal distribution.

Adversarial optimization

We formulated our optimization strategy as follows:

$$\max_{W,p,a} \phi_d(f(\mathbf{W}, p, a))$$

$$\text{s.t. } 0.1 \le w_i \le 0.9$$

$$p \in \{0, 1\}$$

$$a \in \{0, 1\}$$

where f is the CalEnviroScreen algorithm designating tracts as disadvantaged, ϕ_d is a function totalling the number of tracts belonging to a chosen demographic d (for example, political affiliation, race), $\mathbf{W} = \{w_1, \ldots, w_n\}$ is a vector of weights for each variable in the CalEnviroScreen algorithm, p is an indicator variable denoting pre-processing options (percentile-ranking versus z-score standardization), and a is an indicator variable denoting the aggregation methods (multiplication versus averaging). Weight variables were restricted to be between 0.1 and 0.9 to prevent extreme individual weight values. We used the Hooke–Jeeves method to solve the optimization problem[53].

Political affiliation at the tract level was determined by party affiliation in terms of assembly district. For tracts that spanned multiple assembly districts, we attributed those tracts to the districts in which most of their population belonged, in line with how the Climate Change Investments fund attributes tract-level funding to tracts spanning multiple districts. Race was determined by percentage of the population for each tract being of a given race. We calculated the percent change in designated tracts for the party with fewer tracts.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this Article.

Data availability

All data used in this work are publicly available online, and all datasets used are archived at https://doi.org/10.7910/DVN/EVWNC2 (ref. 54). The CalEnviroScreen data are available at https://oehha.ca.gov/calenviroscreen. The CDC PLACES data are available at https://www.cdc.gov/places/index.html. The Climate Change Investments funding dataset is available at https://www.caclimateinvestments.ca.gov/cci-data-dashboard. American Community Survey data are available at https://www.census.gov/programs-surveys/acs/data.html.

Code availability

All code written for this work and a list of packages used are available at https://github.com/etchin/allocativeharm (ref. 55). All analyses were conducted using R (version 4.2.3).

References

1. Mitchell, S., Potash, E., Barocas, S., D'Amour, A. & Lum, K. Algorithmic fairness: choices, assumptions and definitions. *Annu. Rev. Stat. Appl.* **8**, 141–163 (2021).

2. Chen, I. Y. et al. Ethical machine learning in healthcare. *Annu. Rev. Biomed. Data Sci.* **4**, 123–144 (2021).
3. Crawford, K. *The Trouble with Bias* https://nips.cc/virtual/2017/invited-talk/8742#details (2017).
4. Buolamwini, J. & Gebru, T. Gender shades: intersectional accuracy disparities in commercial gender classification. In *Proc. 1st Conference on Fairness, Accountability and Transparency* (eds. Friedler, S. A. & Wilson, C.) 77–91 (PMLR, 2018).
5. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
6. Koenecke, A., Giannella, E., Willer, R. & Goel, S. Popular support for balancing equity and efficiency in resource allocation: a case study in online advertising to increase welfare program awareness. *Proc. International AAAI Conference on Web and Social Media* **17**, 494–506 (2023).
7. Liévanos, R. S. Retooling CalEnviroScreen: cumulative pollution burden and race-based environmental health vulnerabilities in California. *Int. J. Environ. Res. Public Health* **15**, 762 (2018).
8. Flanagan, B. E., Hallisey, E. J., Adams, E. & Lavery, A. Measuring community vulnerability to natural and anthropogenic hazards: the centers for disease control and prevention's social vulnerability index. *J. Environ. Health* **80**, 34–36 (2018).
9. Srivastava, T., Schmidt, H., Sadecki, E. & Kornides, M. L. Disadvantage indices deployed to promote equitable allocation of COVID-19 vaccines in the US. *JAMA Health Forum* **3**, e214501 (2022).
10. Smith, P. C. *Formula Funding of Health Services: Learning from Experience in Some Developed Countries*. Report No. HSS/HSF/DP. 08.1 (World Health Organization, 2008).
11. Suresh, H. & Guttag, J. V. A framework for understanding sources of harm throughout the machine learning life cycle. In *Proc. 1st ACM Conference on Equity and Access in Algorithms, Mechanisms and Optimization* 17 (ACM, 2021); https://doi.org/10.1145/3465416.3483305
12. Kost, R. & Jung, Y. How a tool that tracks California's 'disadvantaged communities' is costing S.F. millions in state funding. *San Francisco Chronicle* (14 December 2021); https://www.sfchronicle.com/bayarea/article/California-has-a-tool-to-fund-its-most-vulnerable-16643115.php
13. Kuruppuarachchi, L. N., Kumar, A. & Franchetti, M. A comparison of major environmental justice screening and mapping tools. *Environ. Manag. Sustain. Dev.* **6**, 59 (2017).
14. Pellow, D. N. Toward a critical environmental justice studies: black lives matter as an environmental justice challenge. *Bois Rev. Soc. Sci. Res. Race.* **13**, 221–236 (2016).
15. Balakrishnan, C., Su, Y., Axelrod, J. & Fu, S. *Screening for Environmental Justice: A Framework for Comparing National, State and Local Data Tools* (Urban Institute, 2022).
16. Mullen, H., Whyte, K. & Holifield, R. Indigenous peoples and the Justice40 screening tool: lessons from EJSCREEN. *Environ. Justice* **16**, 360–369 (2023).
17. Horgan, L. et al. What does Chelsea Creek do for you? A relational approach to environmental justice communication. *Environ. Justice* https://doi.org/10.1089/env.2022.0081 (2023).
18. Dressel, J. & Farid, H. The accuracy, fairness and limits of predicting recidivism. *Sci. Adv.* **4**, eaao5580 (2018).
19. Saisana, M., Saltelli, A. & Tarantola, S. Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *J. R. Stat. Soc. Ser. A Stat. Soc.* **168**, 307–323 (2005).
20. Singh, A., Kempe, D. & Joachims, T. Fairness in ranking under uncertainty. In *Advances in Neural Information Processing Systems* Vol 34 (eds Ranzato, M. et al.) 11896–11908 (Curran Associates, 2021).
21. Kwakkel, J. H., Walker, W. E. & Marchau, V. A. W. J. Classifying and communicating uncertainties in model-based policy analysis. *Int. J. Technol. Policy Manag.* **10**, 299–315 (2010).
22. Walker, W. et al. Comment on 'From data to decisions: processing information, biases and beliefs for improved management of natural resources and environments' by Glynn et al. *Earths Future* **6**, 757–761 (2018).
23. Tsai, H.-J., Wu, P.-Y., Huang, J.-C. & Chen, S.-C. Environmental pollution and chronic kidney disease. *Int. J. Med. Sci.* **18**, 1121–1129 (2021).
24. Boffetta, P. & Nyberg, F. Contribution of environmental factors to cancer risk. *Br. Med. Bull.* **68**, 71–94 (2003).
25. Iqbal, S., Oraka, E., Chew, G. L. & Flanders, W. D. Association between birthplace and current asthma: the role of environment and acculturation. *Am. J. Public Health* **104**, S175–S182 (2014).
26. Tarraf, W., Vega, W. & González, H. M. Emergency department services use among immigrant and non-immigrant groups in the United States. *J. Immigr. Minor. Health* **16**, 595–606 (2014).
27. Angrist, J. D. Lifetime earnings and the Vietnam Era draft lottery: evidence from social security administrative records. *Am. Econ. Rev.* **80**, 313–336 (1990).
28. Stasz, C. & Van Stolk, C. *The Use of Lottery Systems in School Admissions* (Rand Corporation, 2007); https://www.rand.org/pubs/working_papers/WR460.html
29. Yu, B. & Kumbier, K. Veridical data science. *Proc. Natl Acad. Sci. USA* **117**, 3920–3929 (2020).
30. Jeong, Y. & Rothenhäusler, D. Calibrated inference: statistical inference that accounts for both sampling uncertainty and distributional uncertainty. Preprint at https://arxiv.org/pdf/2202.11886.pdf (2022).
31. Vekloff, V. & Abowd, J. *Estimating the Undocumented Population by State for Use in Apportionment* (US Government, 2020); https://www2.census.gov/about/policies/foia/records/2020-census-and-acs/20200327-memo-on-undocumented.pdf
32. Fotheringham, A. S. & Wong, D. W. S. The modifiable areal unit problem in multivariate statistical analysis. *Environ. Plan. Econ. Space* **23**, 933–1086 (1991).
33. Jones, A. P. Indoor air quality and health. *Atmos. Environ.* **33**, 4535–4564 (1999).
34. Nost, E., Horgan, L. & Wylie, S. *Refining CEJST by Including Compliance and Inspection Data and Analysis by Industrial Sector* (Environmental Data & Governance Initiative, 2022); https://envirodatagov.org/wp-content/uploads/2022/04/CEQ-CEJST-Public-Comment-by-EDGI.pdf
35. Kelp, M. M. et al. Data-driven placement of PM2.5 air quality sensors in the United States: an approach to target urban environmental injustice. *GeoHealth* **7**, e2023GH000834 (2023).
36. Nost, E. et al. How gaps and disparities in EPA data undermine climate and environmental justice screening tools (Environmental Data & Governance Initiative, 2022).
37. Lempert, R. J. & Turner, S. Engaging multiple worldviews with quantitative decision support: a robust decision-making demonstration using the lake model. *Risk Anal.* **41**, 845–865 (2021).
38. Walker, W. E. et al. Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integr. Assess.* **4**, 5–17 (2003).
39. Hanna, A., Denton, E., Smart, A. & Smith-Loud, J. Towards a critical race methodology in algorithmic fairness. In *Proc. 2020 Conference on Fairness, Accountability and Transparency* 501–512 (ACM, 2020); https://doi.org/10.1145/3351095.3372826
40. Vera, L. A. et al. When data justice and environmental justice meet: formulating a response to extractive logic through environmental data justice. *Inf. Commun. Soc.* **22**, 1012–1028 (2019).
41. Nozick, R. Distributive justice. *Philos. Public Aff.* **3**, 45–126 (1973).
42. Lamont, J. *Distributive Justice* (Routledge, 2017).

43. Kaswan, A. Distributive justice and the environment. *N. C. Law Rev.* **81**, 1031 (2002).
44. Johansson-Stenman, O. & Konow, J. Fair air: distributive justice and environmental economics. *Environ. Resour. Econ.* **46**, 147–166 (2010).
45. Wang, Y. et al. Air quality policy should quantify effects on disparities. *Science* **381**, 272–274 (2023).
46. Hooker, S. Moving beyond 'algorithmic bias is a data problem'. *Patterns* **2**, 100241 (2021).
47. August, L. et al. CalEnviroScreen 4.0 (OEHHA, 2021); https://oehha.ca.gov/media/downloads/calenviroscreen/report/calenviroscreen40reportf2021.pdf
48. Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R. & Goude, Y. Fast calibrated additive quantile regression. *J. Am. Stat. Assoc.* **116**, 1402–1412 (2021).
49. Imbens, G. & Lemieux, T. Regression discontinuity designs: A guide to practice. *J. Econ.* **142**, 615–635 (2008).
50. Imbens, G. & Kalyanaraman, K. Optimal bandwidth choice for the regression discontinuity estimator. *Rev. Econ. Stud.* **79**, 933–959 (2012).
51. Ho, D. E., Imai, K., King, G. & Stuart, E. A. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.* **15**, 199–236 (2007).
52. Athey, S., Tibshirani, J. & Wager, S. Generalized random forests. *Ann. Stat.* **47**, 1148–1178 (2019).
53. Hooke, R. & Jeeves, T. A. 'Direct Search' solution of numerical and statistical problems. *J. ACM* **8**, 212–229 (1961).
54. Huynh, B. & Chin, E. Replication data for: mitigating allocative tradeoffs and harms in an environmental justice data tool. *Harvard Dataverse* https://doi.org/10.7910/DVN/EVWNC2 (2024).
55. Chin, E. etchin/allocativeharm: Final. *Zenodo*. https://doi.org/10.5281/zenodo.10457118 (2024).

Acknowledgements
We thank A. Yip and E. Knox for helpful discussions. We thank Q. Lin, K. Jen Li and F. Dong for research assistance with the census block group analysis. B.Q.H. acknowledges support by the National Science Foundation Graduate Research Fellowship under grant no. DGE 1656518. The contents of this Article are solely the responsibility of the authors and do not necessarily represent the official views of any agency. Funding sources had no role in the writing of this manuscript or the decision to submit for publication.

Author contributions
B.Q.H. and E.T.C. conceived of the research idea and designed and conducted the primary analyses. B.Q.H. drafted the initial manuscript. E.T.C. created the final figures. A.K. provided technical guidance regarding algorithmic fairness. D.O. performed the census block group analysis. D.E.H. provided technical guidance regarding policy considerations. All authors approved and contributed to the final version of the manuscript.

Competing interests
The authors declare no competing interests.

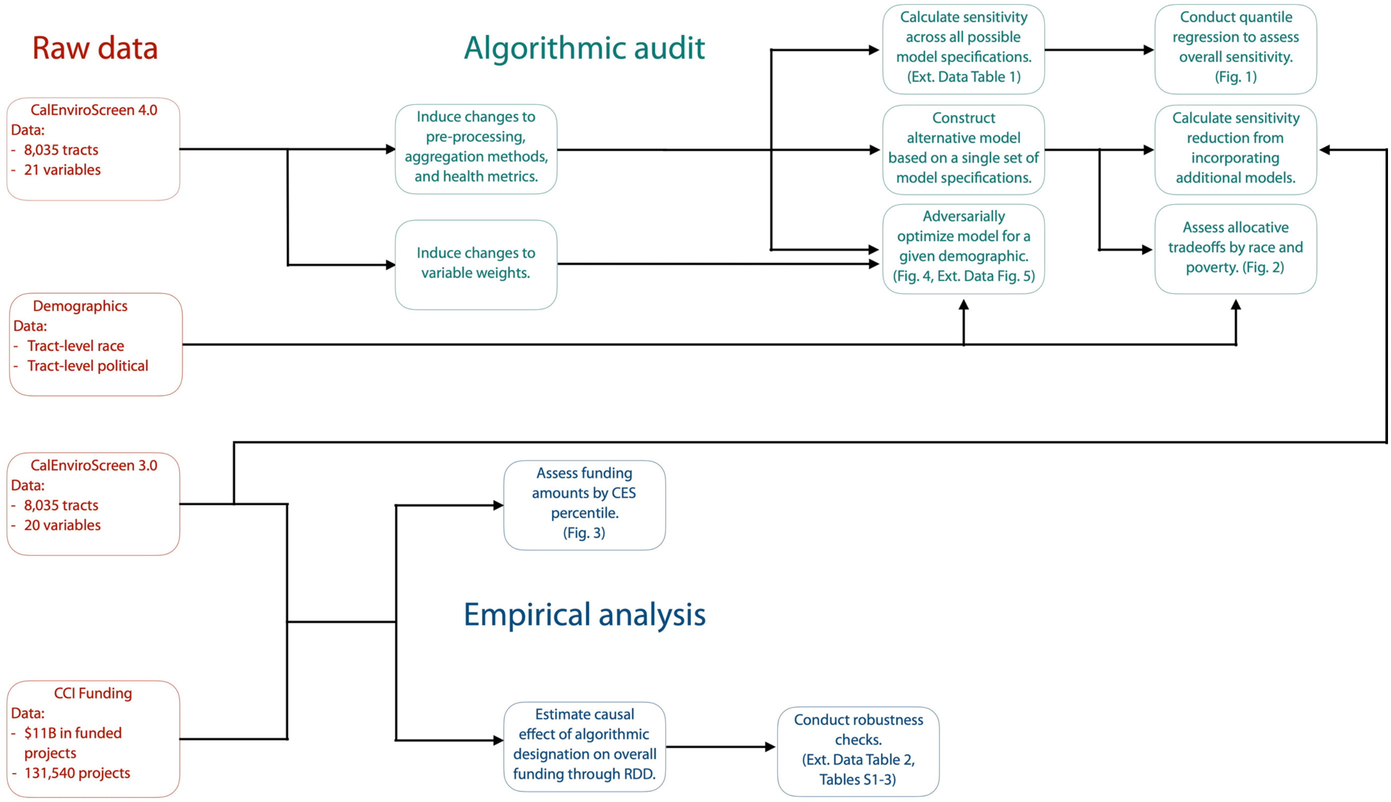
Additional information
Extended data is available for this paper at https://doi.org/10.1038/s42256-024-00793-y.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-024-00793-y.

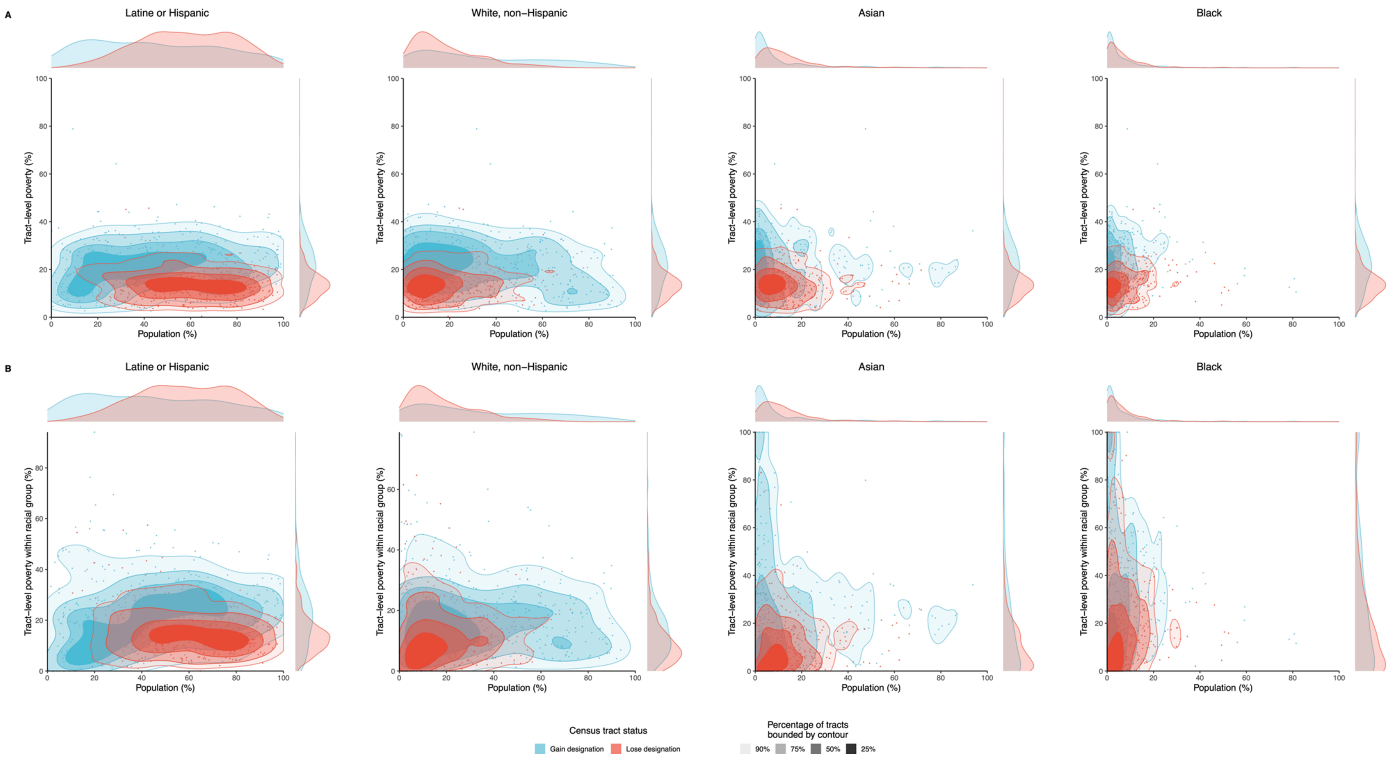
Correspondence and requests for materials should be addressed to Benjamin Q. Huynh.

Peer review information *Nature Machine Intelligence* thanks Lourdes Vera and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

 https://doi.org/10.1038/s42256-024-00793-y

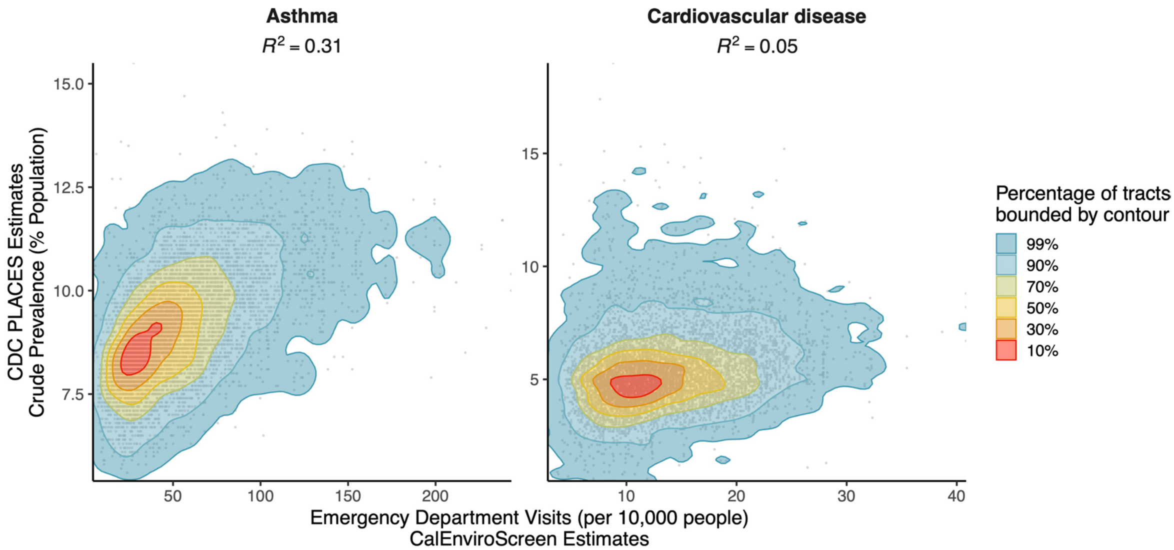


Extended Data Fig. 1 | Analysis pipeline. CES denotes CalEnviroScreen, and RDD denotes regression discontinuity design. Supplementary sensitivity analyses are not depicted (see Supplement).



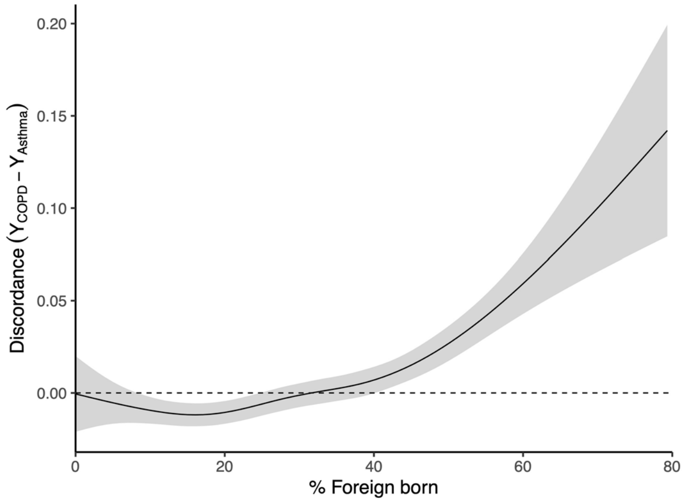
Extended Data Fig. 2 | Allocative tradeoffs between populations in poverty and racially minoritized populations. Comparison of how algorithmically designated tracts are distributed by race and poverty across the current CalEnviroScreen model and an alternative model, among tracts that would change designation status under the alternative model. The alternative model uses a different pre-processing technique, different aggregation technique, and incorporates additional population health variables. Red densities indicate tracts that receive designation under the current model but are not designated under the alternative model. Blue densities indicate tracts gaining designation under the alternative model. Contours are calculated as the smallest regions that bound a given proportion of the data (highest density region). Dots indicate individual tracts. Similar to Fig. 3, except includes data of individual tracts (dots) and varies the y-axis on poverty within racial group and poverty overall.

 https://doi.org/10.1038/s42256-024-00793-y



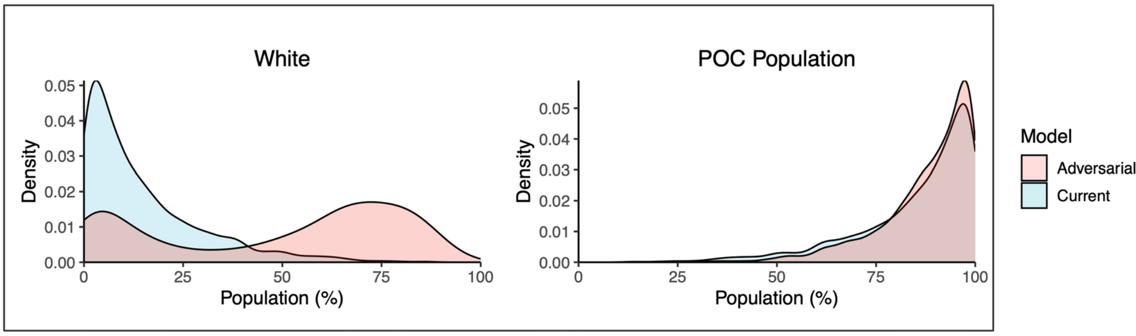
Extended Data Fig. 3 | Low correlations between health indicators used by CalEnviroScreen and an alternative data source. CalEnviroScreen uses tract-level emergency room visits for asthma and heart attacks as health metrics. Comparison data are indicators provided by CDC PLACES: tract-level survey data on history of asthma and coronary heart disease. Contours are calculated as the smallest regions that bound a given proportion of the data (highest density region).

https://doi.org/10.1038/s42256-024-00793-y



Extended Data Fig. 4 | Pairwise discordance between CalEnviroScreen and an alternative model, across tracts with varying levels of foreign-born populations. The alternative model (Y_{COPD}) uses survey data of chronic obstructive pulmonary disease (COPD) as a measure of respiratory health compared to the current CalEnviroScreen model (Y_{Asthma}), which uses emergency room visits for asthma. Higher levels indicate Y_{COPD} designating more tracts as disadvantaged for a given foreign born population percentage. Shaded bars indicate 95% confidence intervals, and black line indicates a smoothing spline from pointwise mean estimates of pairwise discordance. The models are comparable for tracts with fewer than a 30% foreign-born population, suggesting model bias against tracts with high immigrant populations.

 https://doi.org/10.1038/s42256-024-00793-y



Extended Data Fig. 5 | Adversarially optimized distribution of algorithmically designated tracts by race. Blue densities represent designated tracts by the existing CalEnviroScreen model and red densities represent designated tracts by adversarially optimized models, among all algorithmically designated tracts for each model. Plot on the left compares with a model adversarially optimized to increase white populations designated for funding; plot on the right compares with a model adversarially optimized to increase racially minoritized populations designated for funding.

 https://doi.org/10.1038/s42256-024-00793-y

Extended Data Table 1 | Model sensitivity measured as percentage of total tracts that change designation under different model specifications

Pre-processing	Scaling				Percentile-ranking				Overall
Aggregation	Additive		Multiplicative		Additive		Multiplicative		
Additional health metrics	All	None	All	None	All	None	All	None	
% Change in designation	9.8	8.4	7.9	5.4	4.8	1.4	4.3	*Ref.*	16.1

'Scaling' refers to using z-score standardization instead of percentile ranking for pre-processing variables. The 'Additional health metrics' row refers to whether additional population health variables were incorporated. 'Overall' refers to measuring model sensitivity when taking the union of designations from all possible model specifications across scaling, mean, and health.

https://doi.org/10.1038/s42256-024-00793-y

Extended Data Table 2 | Table of effect size estimates from a sharp regression discontinuity approach with varying functional forms, bandwidths, and covariate adjustments

A

Functional form	Bandwidth	Covariate adjustment	Effect size estimate (95% Confidence Interval)
Local Linear	3.86 (IK)	All	103.5 (61.9, 145.1)
		None	103.8 (60.2, 147,3)
	10	All	117.2 (93.2, 141.2)
		None	104.2 (79.3, 129.2)
Quadratic	3.86 (IK)	All	115.8 (47.2, 184.4)
		None	106.8 (34.9, 178.8)
	10	All	101.6 (63.7, 139.6)
		None	92.7 (53.0, 132.4)

B

Functional form	Bandwidth	Covariate adjustment	Effect size estimate (95% Confidence Interval)
Local Linear	3.62 (IK)	All	133.0 (94.2, 171.8)
		None	135.1 (94.9,175.2)
	10	All	167.9 (146.3, 189.4)
		None	152.7 (130.3, 175.0)
Quadratic	3.62 (IK)	All	155.3 (91.6, 219.0)
		None	140.9 (74.8, 207.0)
	10	All	125.9 (92.1, 159.7)
		None	116.9 (81.5, 152.2)

C

Functional form	Bandwidth	Covariate adjustment	Effect size estimate (95% Confidence Interval)
Local Linear	3.37 (IK)	All	100.5 (60.7-140.4)
		None	104.0 (63.0-145.0)
	10	All	122.0 (100.9-143.1)
		None	109.5 (87.6-131.4)
Quadratic	3.37 (IK)	All	99.7 (34.1-165.4)
		None	88.5 (21.0-156.0)
	10	All	93.6 (60.4-126.7)
		None	87.5 (52.9-122.1)

Sensitivity analyses were conducted by varying over model specifications (functional form, bandwidth, and covariate adjustments), as well as across different specifications of the funding dataset: (A) All projects funded from 2017–2021; (B) All projects that explicitly note their funding decisions were based on CalEnviroScreen 3.0.; and (C) All projects funded from 2017–2021 except for projects related to the high-speed rail initiative. Point estimates represent the estimated percent increase in funding resulting from receiving designation. Values in parentheses represent 95% confidence intervals. IK refers to the optimal bandwidths computed using the Imbens-Kalyanaraman algorithm. 'All' refers to adjusting for tract-level race, poverty, and CalEnviroScreen's two primary indicators of pollution burden and population characteristics. Bolded text represents the effect size of the main analysis.

nature portfolio

Corresponding author(s): Benjamin Q Huynh

Last updated by author(s): Jan 4, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No code was used to collect data. All data were downloaded from publicly available websites.
Data analysis	R version 4.2.3 was used to analyze data. We used the following R packages for our analysis: grf (2.2.1) for causal forests, MatchIt (4.5.2) for propensity score matching, qgam (1.3.4) for smooth nonparametric additive quantile regression, rdd (0.57) and rddtools (1.6.0) for regression discontinuity design analyses and Imbens-Kalyanaraman calculations, and optimx (4.30) for Hooke-Jeeves optimization. All code used is publicly available at https://github.com/etchin/allocativeharm.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data used in this work are publicly available online, and all datasets used are archived at https://doi.org/10.7910/DVN/EVWNC2. The CalEnviroScreen data can be found at https://oehha.ca.gov/calenviroscreen. The CDC PLACES data can be found at https://www.cdc.gov/places/index.html. The Climate Change Investments funding dataset can be found at https://www.caclimateinvestments.ca.gov/cci-data-dashboard. American Community Survey data can be found at https://www.census.gov/programs-surveys/acs/data.html.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	No data on sex and gender were collected.
Population characteristics	There were no individual human research participants. We used population-level data from the American Community Survey describing the demographic makeup of each census tract in California in terms of race/ethnicity. We used the definitions for racial and ethnic groups as described by the American Community Survey. We also used publicly available data on health outcomes made available by the state of California, such as number of emergency room visits for cardiac arrest or asthma per census tract. Such information was only available at the population-level.
Recruitment	No participants were recruited for this study. Only publicly available data were used.
Ethics oversight	Because there were no individual human research participants, as well as no usage of protected health information or other confidential information, ethics oversight was not necessary for this study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This was a purely quantitative study, examining the algorithmic outputs of an open-source algorithm using publicly available datasets.
Research sample	The research sample was the publicly available dataset used by the algorithm CalEnviroScreen, consisting of environmental indicators for each census tract in California. We also included demographic information (racial and ethnic breakdown by census tract) from the publicly available American Community Survey. We used the full representative sample (all census tracts in california) to replicate the algorithm.
Sampling strategy	No sampling was performed. We used the entire dataset of all census tracts used by CalEnviroScreen for our study, meaning our "sample" data encompassed the entire population. Therefore, our sample size was sufficient because it encompassed every census tract.
Data collection	All data were downloaded from publicly available online sources, as described in the data availability section of the manuscript.
Timing	All data were downloaded on April 12th 2023.
Data exclusions	No data were excluded from analysis.
Non-participation	No participants dropped out as no individual participants were in the study.
Randomization	Randomization was not applicable to our study -- we used data from all census tracts in California. For our causal analysis, we used a natural experiment approach -- regression discontinuity design -- to assess the difference in funding between those near the decision threshold of CalEnviroScreen's output. The treatment variable was a binary indicator for each tract denoting whether it was designated as disadvantaged by the algorithm. The outcome variable was the log of total funding received per tract. The forcing variable was the CalEnviroScreen percentile rank for each tract. Covariates included the aggregate pollution burden and population characteristics indicators from CalEnviroScreen, and tract-level race and poverty estimates from the American Community Survey.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging