American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

Causal Machine Learning for Fair Allocation of Scarce Public Resources

Dr. Marcus E. Holloway

Associate Professor, University of Melbourne, Australia

Abstract

Public institutions routinely allocate resources whose demand exceeds available supply, including subsidized housing, healthcare interventions, social assistance, educational support, emergency relief, and employment services. Conventional allocation systems often prioritize individuals according to observed need, predicted future risk, waiting time, categorical eligibility, or administrative discretion. Causal machine learning introduces a different analytical possibility by estimating how strongly different individuals or groups are expected to benefit from a particular intervention rather than merely predicting who is currently most disadvantaged or most likely to experience an adverse outcome. Research on heterogeneous treatment effects, causal forests, generalized random forests, and policy learning demonstrates how flexible statistical methods can support treatment targeting under resource constraints. Athey and Wager show that policy learning can optimize treatment assignment from observational data under appropriate causal identification assumptions and application-specific constraints, while Kim and Zubizarreta demonstrate that heterogeneous treatment-effect estimation can itself be conducted under fairness constraints.

This paper develops a governance-oriented framework for applying causal machine learning to scarce public-resource allocation. The proposed approach integrates causal identification, heterogeneous treatment-effect estimation, constrained policy learning, distributional fairness, uncertainty assessment, human oversight, transparency, and contestability. A simulation-based analysis is used to demonstrate the potential tension between outcome efficiency and equity without presenting synthetic values as empirical government results. The simulated fairness–efficiency frontier shows that moderate and high fairness constraints may substantially improve equity while preserving a large proportion of estimated aggregate benefit, whereas extremely restrictive constraints can produce a larger efficiency cost. The study argues that causal targeting should never be interpreted as a purely technical ranking exercise. Allocation legitimacy depends on normative choices concerning whose outcomes count, which inequalities require correction, whether protected attributes may be used to improve substantive fairness, how uncertainty is handled, and which individuals retain meaningful opportunities to challenge automated recommendations. Causal machine learning can contribute to fairer allocation only when causal validity and distributive justice are jointly designed rather than treated as independent objectives.

Keywords: causal machine learning, scarce public resources, heterogeneous treatment effects, policy learning, algorithmic fairness, causal forests, public administration, resource allocation, distributive justice, responsible artificial intelligence

1. Introduction

Scarcity is a persistent feature of public administration. Governments and public-service organizations regularly make decisions about who receives limited housing placements, medical services, income support, educational interventions, rehabilitation programs, emergency assistance, or other publicly financed benefits. Because resources cannot always be provided immediately or universally, administrators require prioritization rules. Traditional systems may rank individuals according to severity of need, waiting time, eligibility categories, predicted risk, professional assessment, or combinations of these criteria. Machine-learning systems have added another possibility by using administrative data to predict outcomes such as hospitalization, homelessness, unemployment, school dropout, or service demand. Yet prediction and intervention are fundamentally different analytical tasks. A person with the highest predicted risk is not necessarily the person who would receive the greatest benefit from the available intervention.

Causal machine learning addresses this distinction by combining causal-inference principles with flexible machine-learning estimators. Wager and Athey developed causal forests to estimate heterogeneous treatment effects, while Athey, Tibshirani, and Wager subsequently generalized the forest framework for broader nonparametric statistical estimation. These approaches allow analysts to investigate whether the effect of an intervention varies according to observed characteristics rather than assuming that a single average treatment effect applies uniformly across a population. The resulting conditional average treatment effect can provide information about who is expected to benefit more or less from a particular resource, subject to the validity of the underlying causal assumptions.

This distinction is particularly important under scarcity. Suppose two applicants face an elevated risk of homelessness. A conventional predictive model may prioritize the applicant with the highest predicted probability of homelessness. A causal allocation model instead asks how much providing a particular housing intervention is expected to change each applicant's outcome relative to the relevant counterfactual. A person at extremely high baseline risk may remain at high risk even after intervention, while another person with moderately lower baseline risk may experience a much larger causal improvement. Athey and Wager's policy-learning framework formalizes the broader problem of learning treatment-assignment rules from observational data while incorporating constraints such as budgets, fairness, or policy simplicity.

However, maximizing estimated causal benefit does not automatically produce a fair public policy. Research on contextual resource allocation demonstrates that fairness in allocations and fairness in outcomes can conflict and that vulnerability-based prioritization does not necessarily generate equal outcomes across social groups. Nabi, Malinsky, and Shpitser further show that causal pathways can be incorporated directly into definitions of fair policies, emphasizing that historical discrimination can influence both observed data and policy consequences. Kim and Zubizarreta likewise demonstrate that heterogeneous treatment effects can be estimated under explicit fairness constraints, thereby connecting causal learning with distributional objectives rather than treating fairness as an external post-processing exercise.

The present study therefore examines causal machine learning not simply as an optimization technology but as an institutional framework for allocating socially significant goods. It proposes a methodological architecture combining identification, causal-effect heterogeneity, policy optimization, explicit resource constraints, fairness constraints, uncertainty, transparency, and human review. Because no administrative dataset or completed field experiment was supplied, the paper uses explicitly simulated analytical values. The study does not claim that the reported scores represent the performance of any real government program. Its contribution is methodological: it demonstrates how public institutions can move from prediction-centered allocation toward causal and fairness-aware policy learning while preserving the normative accountability that legitimate public-resource allocation requires.

2. Objectives of the Study

2.1 To Distinguish Predictive Prioritization from Causal Allocation

The first objective is to clarify the difference between identifying individuals who are likely to experience adverse outcomes and identifying individuals for whom a particular public intervention is likely to produce the largest causal improvement. This distinction is essential because predictive risk models estimate expected outcomes conditional on observed characteristics, whereas causal models attempt to estimate differences between potential outcomes under alternative intervention states.The study therefore evaluates heterogeneous treatment-effect estimation as a basis for resource targeting and explains why prediction alone may be insufficient for policy optimization.

2.2 To Integrate Fairness into Causal Policy Learning

The second objective is to develop a framework in which fairness is incorporated during allocation-policy construction rather than assessed only after an optimized policy has been generated.The proposed approach considers multiple fairness concepts, including equality of access, allocation parity, outcome equity, priority for disadvantage, minimum-benefit guarantees, and protection against impermissible causal pathways. It recognizes that these goals may conflict and therefore require explicit public-policy judgment rather than silent optimization.

2.3 To Develop a Governance Model for Responsible Public-Sector Deployment

The third objective is to establish practical safeguards for institutions considering causal machine learning in high-stakes resource allocation.The proposed governance model addresses causal identification, data quality, model uncertainty, fairness specification, public justification, human review, monitoring, appeal mechanisms, and periodic re-evaluation. The objective is not autonomous allocation but decision support that remains institutionally accountable.

3. Review of Literature

3.1 Heterogeneous Treatment Effects and Causal Machine Learning

Traditional program evaluation frequently begins with the average treatment effect, which asks whether an intervention improves outcomes on average. Yet public programs can generate heterogeneous effects because individuals differ in needs, environments, barriers, and responsiveness.

These developments are valuable for resource allocation because public authorities rarely face homogeneous populations. Employment programs may be especially beneficial for particular groups of job seekers, a preventive health intervention may yield larger gains for certain patients, and a housing intervention may produce different stability outcomes according to household circumstances.

Yet treatment-effect heterogeneity is not synonymous with individual destiny. CATE values describe conditional average effects among individuals sharing measured characteristics; they should not automatically be interpreted as perfectly known individual treatment effects. Moreover, causal estimates remain dependent on identification assumptions, data quality, overlap, model specification, and the policy environment in which historical treatments were assigned.

3.2 Policy Learning Under Resource Constraints

Estimating heterogeneous effects is only one stage of the allocation problem. Public institutions must translate estimated effects into operational rules subject to budget, capacity, legal, and administrative constraints.

Tang, Koçyiğit, Rice, and Vayanos study online allocation of scarce societal resources including housing-related applications and develop policies that incorporate resource constraints and fairness requirements. Their work demonstrates the possibility of directly integrating fairness into allocation rules rather than assuming efficiency maximization alone defines the appropriate public objective.

Mate, Wilder, Taneja, and Tambe examine randomized trials for evaluating algorithmic resource-allocation policies, highlighting the methodological importance of rigorous policy evaluation rather than assuming an optimized rule will necessarily improve deployed outcomes. This point is crucial because an apparently superior policy estimated retrospectively can perform differently when introduced into a dynamic service environment.

3.3 Fairness, Causality, and Distributional Justice

Algorithmic fairness becomes especially complex when decisions concern scarce goods. Equal allocation rates, equal outcomes, need-based priority, causal-benefit maximization, and priority for historically disadvantaged groups may all represent plausible fairness objectives, but they are not equivalent.

Jo and colleagues' work on contextual resource allocation demonstrates that allocation fairness and outcome fairness are frequently incompatible. Their analysis also shows that even vulnerability scores that are perfectly calibrated can generate unequal outcomes across groups. This finding has major implications for public administration because apparent neutrality in the allocation rule does not guarantee equality in downstream consequences.

Nabi, Malinsky, and Shpitser use causal inference to formalize fairness constraints around impermissible causal pathways. Their approach recognizes that protected characteristics may affect outcomes through historically structured pathways and that policy design requires an explicit judgment about which causal mechanisms should or should not influence allocation.

Kim and Zubizarreta develop a framework for heterogeneous treatment-effect estimation under fairness constraints and characterize trade-offs between fairness and statistical performance. Rehill and Biddle further argue that fairness concerns in causal machine learning differ from standard predictive-fairness settings because causal models often provide information to policymakers rather than directly determining final decisions. They emphasize the importance of ensuring that decision-makers receive sufficiently transparent information to make justified normative judgments.

Scarcity may also create situations in which deterministic ranking itself is ethically questionable. Jain, Creel, and Wilson argue that machine-learning-supported scarce-resource allocation may sometimes require randomization because similarly deserving individuals can possess comparable claims to a resource even where estimated scores differ slightly. This insight suggests that fairness-aware causal allocation should preserve lotteries or randomized tie-breaking where uncertainty is high and claims are substantively similar.

4. Research Methodology

4.1 Research Design and Evidence Base

This study adopts a conceptual-methodological design supported by simulation-based policy analysis. It does not use real applicant records, confidential administrative data, or completed government allocation decisions.

The methodological synthesis draws on causal forests, generalized random forests, heterogeneous treatment-effect estimation, policy learning, fairness-constrained policy optimization, and responsible-AI governance. Athey and Wager's work provides the core policy-learning logic, while later fairness-oriented research demonstrates how equity considerations can be incorporated directly into heterogeneous effect estimation and policy design.

The analytical procedure consists of three conceptual stages. First, a causal estimand is defined for the relevant public intervention. Second, heterogeneous treatment effects are estimated conceptually and translated into a constrained allocation policy. Third, fairness constraints are introduced at increasing strengths to examine their effect on simulated aggregate efficiency and equity.
No numerical result is presented as an observed population parameter.

4.2 Causal Identification and Policy-Learning Procedure

Before causal machine learning can be applied, the public agency must specify a credible identification strategy. Where observational administrative data are used, relevant assumptions may include consistency, appropriate adjustment for confounding, overlap or positivity, and a defensible treatment and outcome definition. Policy-learning methods can also operate under other identification strategies, but the causal assumptions must be explicitly connected to the institutional setting. Athey and Wager's policy-learning framework accommodates different identification strategies rather than assuming predictive association is sufficient.

The proposed analytical sequence is:

[X \rightarrow \text{Causal Identification} \rightarrow \hat{\tau}(X) \rightarrow \text{Policy Optimization} \rightarrow \text{Fairness Constraint} \rightarrow \text{Human Review}]

A fairness-constrained optimization problem can be represented conceptually as:

[\max_{\pi} E[\pi(X)\hat{\tau}(X)]]

subject to:

[E[c(X)\pi(X)] \leq B]

and:

[F(\pi)\leq \delta]
where (F(\pi)) represents a selected fairness disparity and (\delta) represents the maximum tolerated disparity.
The fairness constraint cannot be chosen solely by the data scientist because its interpretation is inherently normative.

4.3 Evaluation Framework

Table 1: Methodological Framework for Fair Causal Resource Allocation

Stage	Principal Question	Technical Requirement	Fairness Requirement	Governance Control
Problem Definition	What public outcome should improve?	Explicit causal estimand	Outcome must not systematically devalue protected groups	Public-purpose justification
Data Construction	Are historical records appropriate?	Data provenance and confounder assessment	Identify historical and measurement bias	Data audit
Causal Identification	Can treatment effects be identified?	Exchangeability/identification strategy, consistency and overlap	Check differential overlap and uncertainty	Independent methodological review
HTE Estimation	Who is expected to benefit?	Validated CATE estimation	Assess heterogeneous errors across groups	Subgroup diagnostics
Policy Learning	Who receives limited resources?	Budget- or capacity-constrained optimization	Explicit fairness constraint	Transparent policy rule

Source: Author-developed methodological synthesis based on causal policy-learning and fairness literature.
The framework explicitly distinguishes causal validity, allocation efficiency, and normative fairness. A model can perform strongly in one dimension while failing in another.

5. Analysis

5.1 Comparison of Alternative Allocation Logics

Public resource allocation can be structured according to different decision principles. To clarify the distinctive role of causal machine learning, the present study compares five illustrative policy logics.

Table 2: Comparative Assessment of Scarce-Resource Allocation Policies

Allocation Policy	Primary Criterion	Main Strength	Principal Fairness Risk	Appropriate Use
Waiting-Time Rule	Time already waiting	Simple and transparent	Ignores urgency and differential benefit	Comparable claims with limited clinical/social differentiation
Vulnerability Rule	Severity of baseline need	Prioritizes disadvantage	Highest need may not equal highest intervention benefit	Rights- or need-based public programs
Predictive-Risk Rule	Probability of adverse outcome	Can identify high-risk cases	Confuses risk with causal responsiveness	Screening and preventive monitoring
Causal-Benefit Rule	Estimated treatment effect	Targets intervention where expected impact is largest	May disadvantage groups with historically lower measured responsiveness	Outcome-oriented policy under strong causal validity
Fairness-Constrained Causal Policy	Causal benefit plus explicit equity constraint	Balances impact and distributive objectives	Requires normative choice of fairness criterion	High-stakes public allocation with transparent governance

The comparison reveals why causal machine learning should not simply replace existing public criteria. A vulnerability-based rule and a causal-benefit rule answer different moral questions.

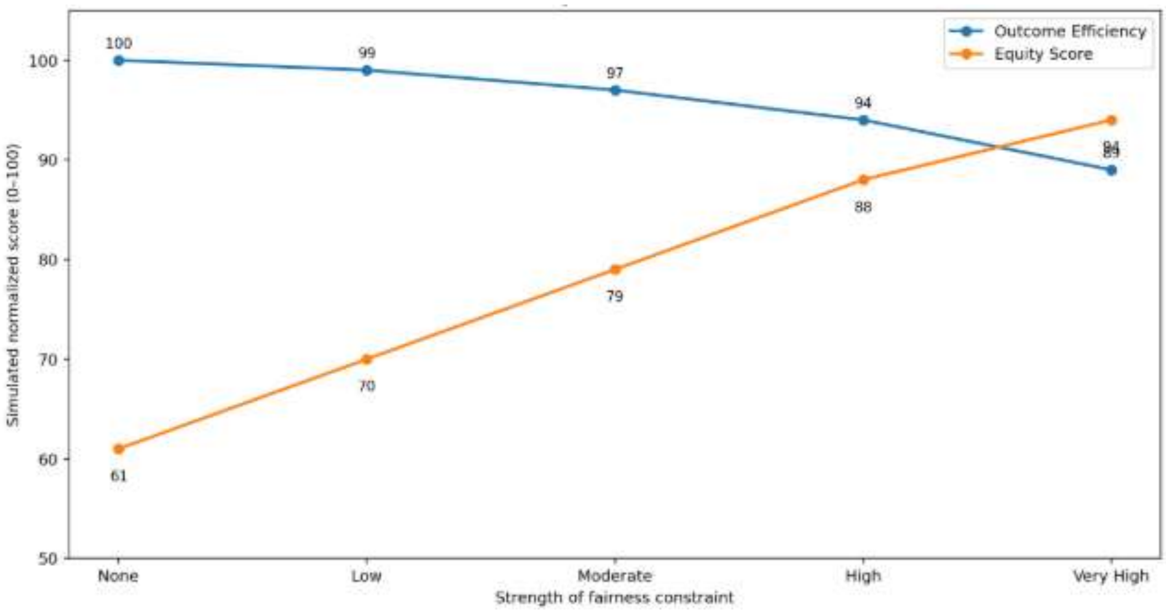
5.2 Simulation of the Fairness–Efficiency Trade-Off

The simulation assumes an unconstrained causal-benefit policy normalized to an aggregate outcome-efficiency score of 100. Fairness constraints are progressively strengthened while a separate equity index is increased from 61 to 94.

The values are hypothetical methodological examples.

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)
E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

Graph 1: Simulated Fairness–Efficiency Frontier in Causal Resource Allocation



Source: Author-generated simulation. No real public-sector beneficiaries or administrative records were used.

61. Introducing a low fairness constraint raises equity to 70 while reducing estimated efficiency only to
99. A moderate constraint produces an equity score of 79 while retaining 97% of normalized efficiency. Under the high-fairness scenario, equity reaches 88 while efficiency remains at
94. Only the very-high constraint produces a more substantial decline in the efficiency index, reaching 89 while equity rises to 94.

This pattern is illustrative but conceptually consistent with fairness-constrained policy-learning research showing that fairness and aggregate policy performance can sometimes be traded off rather than existing as an all-or-nothing choice. The key result is therefore not that a particular fairness parameter is universally optimal. The appropriate constraint depends on the public value being allocated, the legal context, the severity of deprivation, the reliability of treatment-effect estimates, and society's normative commitments.

5.3 Governance Interpretation of Causal Allocation

The analysis suggests that causal machine learning is most defensible when used as a structured decision-support mechanism rather than an autonomous entitlement system.

First, the intervention must have a well-defined causal objective. Public administrators should not optimize an outcome merely because it is easy to measure. A housing system, for example, might optimize short-term program exit while neglecting long-term stability.

Second, uncertainty should affect allocation decisions. Individuals with estimated treatment effects of 0.38 and 0.37 should not necessarily be treated as substantively different when confidence intervals are wide.

Third, historical allocation patterns can affect identification. If particular groups rarely received a service, causal effects for those groups may be weakly identified. A data-driven system should not convert absence of historical evidence into evidence of low benefit.

Fourth, fairness constraints should be publicly interpretable. Authorities should state whether the objective concerns equal access, equal outcomes, priority to disadvantage, minimum benefit guarantees, or another distributive principle.

Fifth, allocation systems should preserve procedural safeguards. Individuals should be able to correct erroneous data and challenge determinations where algorithmic information materially affects access to essential resources.

6. Discussion

6.1 From Predicting Need to Estimating the Consequences of Intervention

The most important conceptual contribution of causal machine learning to public-resource allocation is the separation of prediction from intervention. Predictive machine learning estimates what is likely to happen given observed patterns, whereas causal analysis seeks to determine how outcomes would differ under alternative actions. This difference matters because scarce resources are interventions rather than labels. A public authority does not merely need to know which household is likely to remain homeless, which student is likely to fail, or which patient is likely to deteriorate. It must also determine whether providing a particular resource changes that outcome and how that change differs across eligible recipients. Causal forests and related heterogeneous treatment-effect methods provide one technical route for estimating such differences, while policy-learning frameworks translate them into feasible assignment rules.

Nevertheless, causal estimation does not remove the ethical foundations of public decision-making. Maximizing treatment effects can privilege responsiveness rather than need. If an intervention produces modest measurable gains for individuals facing severe disadvantage but larger gains for those starting from better conditions, a pure benefit-maximization policy could systematically deprioritize the most vulnerable. This is not necessarily a statistical error; it may be the mathematically correct consequence of an ethically incomplete objective. Public agencies should therefore determine whether causal benefit is the primary distributive principle, one factor among several, or a secondary criterion applied only after minimum need thresholds have been satisfied. The causal model supplies information about consequences, while democratic and administrative institutions remain responsible for determining which consequences should govern allocation.

6.2 Fairness Cannot Be Reduced to a Single Mathematical Constraint

Fairness-aware causal policy learning creates an important improvement over unconstrained optimization because it allows allocation objectives to incorporate distributional considerations explicitly. However, the literature demonstrates that different fairness criteria can conflict. Jo and colleagues show that equality in allocation and equality in outcomes may be incompatible in contextual resource-allocation settings. Nabi and colleagues further demonstrate that causal definitions of fairness depend on judgments regarding which pathways from protected characteristics to outcomes are permissible. As a result, the selection of a fairness metric is not merely a technical hyperparameter decision. It represents a substantive public-policy position concerning equality, responsibility, historical disadvantage, and legitimate differentiation.

The simulation in this paper illustrates this point through a fairness–efficiency frontier. Moderate constraints substantially improve the synthetic equity index with only limited loss of normalized aggregate outcome efficiency, whereas more restrictive constraints eventually produce larger efficiency costs. The existence of such a frontier does not establish which point should be selected. A public-health emergency, housing program, educational grant system, and employment service may legitimately select different positions because the moral significance of unequal allocation and unequal outcomes differs across institutional contexts. Transparent governance requires public agencies to state what fairness objective was selected, why that objective is relevant to the resource, which competing objectives were considered, and how the consequences of alternative choices were evaluated.

Fairness assessment must also account for uncertainty. A deterministic system that ranks people according to tiny differences in estimated treatment effect can create an appearance of scientific precision that the data do not support. Randomized allocation among similarly situated individuals may therefore be more defensible than deterministic ordering when estimated differences fall within meaningful uncertainty ranges. Jain, Creel, and Wilson's argument for randomization in machine-learning-supported scarce-resource allocation provides a useful reminder that equal moral claims cannot always be reduced to an infinitesimal difference in a model score.

6.3 Institutional Accountability, Human Review, and Policy Evaluation

Causal machine learning should be governed as part of a public institution rather than as an isolated predictive artifact. NIST's AI Risk Management Framework emphasizes organizational processes for mapping, measuring, managing, and governing AI risks, including explicit assessment of fairness and bias. In scarce-resource allocation, this institutional perspective requires clear responsibility for the causal assumptions, policy objective, fairness criterion, data sources, model updates, decision authority, appeals, and monitoring arrangements. A technically sophisticated estimator does not eliminate the need for an accountable official who can explain why the resulting policy is legitimate.

Human review should also be structured around the distinction between model information and final allocation judgment. Rehill and Biddle argue that causal machine learning in policymaking frequently informs human decision-makers rather than replacing them, which makes transparency regarding model limitations particularly important. Human reviewers should therefore receive not merely a ranking but information concerning estimated treatment effects, uncertainty, data limitations, relevant fairness constraints, and the reasons an individual was prioritized or deprioritized. The reviewer should have genuine authority to respond to exceptional circumstances rather than simply approving the algorithmic result.

Finally, deployment should be treated as an evaluable intervention. Mate and colleagues emphasize rigorous randomized evaluation of algorithmic resource-allocation policies. Public institutions should measure whether an allocation system actually improves outcomes after introduction, whether benefits remain equitably distributed, whether the population changes, and whether the model alters behavior or data collection in ways that invalidate historical assumptions. A causal model can become unreliable when the policy environment it helped create differs materially from the environment in which it was estimated. Continuous evaluation is therefore part of causal validity rather than an optional governance addition.

7. Conclusion

Causal machine learning offers an important conceptual advance for the allocation of scarce public resources because it directs attention from predicting adverse outcomes toward estimating the consequences of intervention. In settings where governments must decide how to allocate limited housing placements, health interventions, educational services, welfare programs, or other socially valuable resources, heterogeneous treatment-effect estimation can help identify situations in which the same intervention produces different expected benefits across recipients. Policy-learning methods can subsequently translate these estimates into feasible allocation rules subject to capacity, cost, and administrative constraints. This creates the possibility of designing allocation systems around expected causal impact rather than relying exclusively on historical practice, predicted risk, or generalized assumptions about average program effectiveness.

The study nevertheless demonstrates that causal efficiency cannot serve as a complete theory of public fairness. A policy that maximizes total expected benefit may distribute resources differently from a policy that prioritizes the worst off, equalizes access, reduces group disparities, or guarantees minimum protection for vulnerable populations. Fairness-aware causal machine learning is therefore valuable precisely because it makes some of these tensions explicit. The simulated fairness–efficiency analysis illustrates that stronger equity requirements can improve distributional performance while preserving substantial aggregate efficiency, although the particular trade-off will vary across applications. No universally correct fairness parameter follows from the mathematics because the choice ultimately depends on the social purpose of the program, rights of affected populations, historical inequality, institutional mandate, and democratic legitimacy.

Responsible implementation requires substantially more than fitting a causal forest and ranking individuals according to estimated treatment effects. Public agencies must establish a credible causal identification strategy, examine differential uncertainty and overlap, validate outcome definitions, document data provenance, assess historical discrimination, specify resource constraints, articulate fairness objectives, provide meaningful human oversight, and create accessible mechanisms for correction and appeal. Particular caution is required when groups are poorly represented in historical treatment data because weak evidence about treatment effectiveness should not automatically be transformed into reduced eligibility. Where individuals possess substantially equivalent claims and model uncertainty makes distinctions unstable, randomized tie-breaking or lottery mechanisms may be more legitimate than deterministic ranking.

The central conclusion is that causal machine learning can inform fair allocation, but it cannot independently define what fairness requires. Its strongest contribution is the ability to provide better information about the likely consequences of alternative interventions. The authority to determine which outcomes matter, which inequalities deserve priority, which causal pathways are legitimate, and how scarce public goods should ultimately be distributed remains a responsibility of accountable public institutions. Future empirical research should validate fairness-constrained causal policies using real administrative datasets, prospective trials, sensitivity analyses, stakeholder participation, and longitudinal monitoring. When embedded within such governance arrangements, causal machine learning can become a valuable instrument for evidence-informed public administration without allowing optimization to replace distributive justice.

References

1. Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge University Press.
2. Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.
3. Hernán, M. A., & Robins, J. M. (2020). *Causal Inference: What If*. Chapman & Hall/CRC.
4. Morgan, S. L., & Winship, C. (2015). *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (2nd ed.). Cambridge University Press.
5. Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences, 113*(27), 7353–7360. https://doi.org/10.1073/pnas.1510489113
6. Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics, 47*(2), 1148–1178. https://doi.org/10.1214/18-AOS1709
7. Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association, 113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839
8. Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. https://doi.org/10.1111/ectj.12097
9. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.
10. Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems, 30*, 4066–4076.
11. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, *29*, 3315–3323.
12. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, 43:1–43:23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43
13. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047
14. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*(3), 671–732.
15. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*.
16. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68). ACM. https://doi.org/10.1145/3287560.3287598

17. Kleinberg, J., Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, *105*(5), 491–495. https://doi.org/10.1257/aer.p20151023
18. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342
19. Kasy, M., & Abebe, R. (2021). Fairness, equality, and power in algorithmic decision-making. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 576–586.
20. Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106. https://doi.org/10.1257/jep.31.2.87