frontiers | Frontiers in Behavioral Economics

TYPE Opinion
PUBLISHED 03 December 2024
DOI 10.3389/frbhe.2024.1503793

OPEN ACCESS

EDITED BY
Armenak Antinyan,
Cardiff University, United Kingdom

REVIEWED BY
Miloš Fišar,
Masaryk University, Czechia

*CORRESPONDENCE
Sanchayan Banerjee
✉ sanchayan.1.banerjee@kcl.ac.uk

RECEIVED 29 September 2024
ACCEPTED 14 November 2024
PUBLISHED 03 December 2024

CITATION
Banerjee S and Veltri GA (2024) Harnessing pluralism in behavioral public policy requires insights from computational social science. *Front. Behav. Econ.* 3:1503793. doi: 10.3389/frbhe.2024.1503793

Harnessing pluralism in behavioral public policy requires insights from computational social science

Sanchayan Banerjee[1,2]* and Giuseppe A. Veltri[3]

[1]The Policy Institute, King's College London, London, United Kingdom, [2]Institute for Environmental Studies, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, [3]Department of Sociology and Social Research, University of Trento, Trento, Italy

KEYWORDS

behavioral public policy, computational social science, large language models, heterogeneity, mechanisms, personalization

Introduction

Behavioral public policy has increasingly proposed alternate toolkits, such as the boost or nudge+, to counter several limitations of nudges. While there is growing evidence that these toolkits can be more effective and legitimate than nudging in certain contexts (Banerjee et al., 2023), practitioners and policymakers struggle to differentiate between them. In this correspondence, we argue that an increasing pluralism through multiple toolkits in the toolbox of the behavioral scientist is meaningful if and only if there are clearer rules to uniquely design and identify these policies. In fact, to harness pluralism, practitioners must have guidance and thumb rules that help them choose one toolkit over the other in a given scenario. We suggest this is possible with recent developments in computational social science, using insights from machine learning and large language models (LLMs), that can be used to understand heterogeneity in the effectiveness of behavioral interventions (BIs). For example, LLMs can enhance experimental designs by generating synthetic participants to complement real ones, improving covariate balancing, model validation, and adaptive randomization, ultimately making studies more efficient and robust in exploring treatment effect heterogeneity. Uncovering this heterogeneity can establish causal evidence on the precursors and mechanisms underlying behavioral change strategies, enabling researchers to personalize these BI toolkits.

The current focus of behavioral research is highly applied and interdisciplinary, leading to a blurring of conceptual boundaries between different behavioral techniques (see Table 1 below for an overview). This trend has made it challenging to distinguish between various toolkits, turning them into mere labels. For instance, the concept of nudging, originally centered on reflexive psychological prompts like defaults, has expanded to include more thoughtful interventions like educative nudges, overlapping with other techniques like boosting, which aim to enhance human capabilities. This overlap is evident in how educative nudges and short-term boosts both address specific problems in a similar manner (Hertwig and Grüne-Yanoff, 2017, p. 397). Additionally, nudging and boosting share similarities with "thinks" (John et al., 2009), which are large-scale educational policies designed to improve societal decision-making. Similarly, nudge+ interventions, combining nudging with reflective prompts, also overlap with system-2 nudges, as they both promote deliberative responses to behavioral problems. Recently, scholars have argued that nudge+ and boosting both enhance agency, clubbing them together within a behavioral agency framework (Banerjee et al., 2024). The lack of clear distinctions between these toolkits has made them less useful for practitioners and potentially more confusing to navigate.

TABLE 1 Pluralism in behavioral public policy toolkits.

Behavioral toolkit	Definition	Policy mechanism	Cognitive mechanism	Target
Nudge	Nudge, as defined by Thaler and Sunstein (2008, p. 8), refers to "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid"	Reducing friction of making "optimal" or "rational" choices. Over time, nudges have expanded to include informational campaigns or educational programs that make people think or educate them to be better versions of themselves, overlapping with boosts or nudge+ interventions	Bypassing one's biases working through automatic (type-1) processes but has grown to include more reflective (type-2) processes. Follows Heuristics and Biases paradigm	Ends (final behaviors)
Boost	Boost, as defined by Hertwig and Grüne-Yanoff (2017, p. 977), refers to "as interventions that target competences rather than immediate behavior (Table 1). The targeted competences can be specific to a single domain…or generalize across domains (e.g., statistical literacy)"	Building and fostering better competencies and skillsets so people can use these new rules to make "optimal" and "rational" decisions. While not generalisable, some (short-term) boosts are merely educational policies which teach people new rules or heuristics, overlapping with reflective nudges	Creating better heuristics for people. Typically starts with reflective processes (type-2) which when repeated results in new competencies that become automatic (type-1)	Means (competences)
Nudge+	Nudge+, as defined by Banerjee and John (2024, p. 69), refers to "a modification of the toolkit of behavioral public policy [which] incorporates an element of reflection—the plus—into the delivery of a nudge, either blended in or made proximate." The plus can be delivered before, after or alongside the nudge	A prompt beside a nudge that encourages people to think if the said nudge is a good fit for themselves. Often thought of a knee-jerk reaction, a quick check, that can de-anchor people from the nudge and enable them to evaluate it objectively. Overlaps with system-2 or educative nudges	A hybrid toolkit that taps into reflexive (type-1) and reflective (type-2) processes at the same time. If the cue generated from the prompt ("the plus") does not align with the response under nudge, type-2 processes kick in	The nudge in question and its objective (ends as such)

Despite this, practitioners have extensively utilized pluralism in behavioral public policy by relying on different BIs. For instance, behavioral "nudge units" have implemented approaches that are predominantly pluralistic rather than strictly labeled—the MINDSPACE report, an early classification attempt in behavioral public policy, assembled a variety of BIs that went beyond simple nudges. Scholars advocating for these alternative toolkits argue that conceptual clarification aids in comparative analysis. However, why hasn't this been particularly useful thus far?

One explanation is that efforts to delineate BIs and establish clear guidelines have been hampered by limitations in data and methodology for inferring causal mechanisms and establishing precursors. For instance, factors like motivation and conscientiousness have been proposed as necessary prerequisites for boosting people's competencies (Hertwig, 2017). However, determining whether individuals meet these criteria requires identifying specific groups either through behavioral profiling before intervention delivery or reassessing them afterward based on their response to treatment—both of which are challenging tasks. In general, understanding heterogeneity and meeting prerequisites for effectively administering BIs has remained a methodological challenge related to causally identifying diverse treatment effects. This difficulty is amplified by the lack of a clear conceptual framework for tailoring interventions to individuals, as well as the inability to causally infer tailoring effects, even when attempted, due to self-selection bias leading to violations of experimental conditions (stable unit treatment value assumption).

In this article, we outline new advancements in computational social science methods, including large language models, and discuss how these methods can be used in conjunction with conventional BPP toolkits in the pursuit to uncover heterogeneity in effect sizes and thereby understand mechanisms underlying behavior change. We propose that harnessing pluralism in the toolkit of the policymaker is possible using these recent developments in computational methods which in turn is key to personalizing the delivery of behavioral interventions.

Advances in computational social sciences for BPP

Recent progress in causal inference methods, particularly with the integration of machine learning algorithms within computational social science, has marked a significant advancement. We now possess various techniques for estimating the heterogeneous treatment effects using causal machine learning, applicable to both experimental and observational data. From causal trees and forests to metalearners and Bayesian statistics approaches (see Table 2 for an overview of their features), these methodologies have transformed our comprehension of the subtleties in treatment effects across different subpopulations, enabling us to identify precursors and mechanisms of behavior change.

Causal trees and forests, pioneered by Athey and Imbens (2016), utilize the hierarchical structure of decision trees to divide data into subgroups with distinct treatment effects. By iteratively splitting the data based on interacting covariates, these methods pinpoint the most relevant variations in treatment effects. For example, a causal tree might first split the data by age group and then by cognitive reflection capacities (using CRTs) within each age group, highlighting how treatment effects differ across these dimensions. The resulting tree or ensemble of trees offers an

TABLE 2 Overview of computational methods for heterogeneity analysis.

FeatureS	Causal trees and forests	Meta learners (X-learner, R-learner)	Bayesian techniques (BART, BCF)
Model type	Ensemble of decision trees	Combination of multiple models	Bayesian ensemble methods
Approach	Non-parametric, tree-based	Flexible (depends on base learners)	Probabilistic, Bayesian inference
Handling of non-linearity	Strong	Depends on base learners	Strong
Handling of interactions	Automatic detection	Depends on base learners	Captures complex interactions
Interpretability	High (tree structures are interpretable)	Varies (depends on base learners)	Moderate (through posterior analysis)
Data efficiency	High	Moderate	Requires larger samples
Computational intensity	High	Moderate	High
Scalability to high dimensions	Good	Depends on base learners	Moderate
Direct optimization for heterogeneity	Yes	Indirect	Yes
Separate models for treatment/control	No	Yes	No
Flexibility in model choice	Fixed (tree-based methods)	High (any suitable base learner)	Fixed (Bayesian tree models)

understandable and data-driven approach to exploring treatment effect diversity across groups of people varying in their reflective potential. Causal trees and forests are particularly useful when the treatment effect varies based on observable characteristics, as they can identify the specific subgroups that benefit most from the treatment. If, for example, a certain age group with a higher reflective potential is associated with a larger uptake of the treatment, this localized effect could suggest such groups of people might be more amenable to reflective BIs such as nudge+ vs. a nudge.

On the other hand, meta learners like the X-learner (Künzel et al., 2019) and the R-learner (Nie and Wager, 2021) adopt a different strategy, combining multiple machine learning models to estimate the conditional average treatment effect (CATE). Typically, these methods entail training separate models for the treatment and control groups, then merging their predictions to estimate the CATE. For instance, using the same example as above, the X-learner would train a model on the treated units and another on the control units within each age group and CRT levels, using the difference between their predictions to estimate the treatment effect for each subgroup. Leveraging the strengths of diverse machine learning algorithms, meta learners can yield more precise and resilient estimates of treatment effect variability. Meta learners are advantageous when the relationship between the covariates and the treatment effect is complex and cannot be easily captured by a single model.

Furthermore, Bayesian statistical techniques such as Bayesian Additive Regression Trees (BART) (Hill, 2011) and Bayesian Causal Forests (BCF) (Hahn et al., 2020) provide a probabilistic framework for estimating heterogeneous treatment effects. These methods integrate prior knowledge and uncertainty quantification into the estimation process. By sampling from the posterior distribution of treatment effects, Bayesian approaches furnish point estimates and credible intervals that gauge the uncertainty surrounding the estimated treatment effects. For example, BART can analyse treatment effects within the same subgroups of age and CRT levels, generating a range of possible treatment effects, giving researchers a sense of how confident they can be in the estimates and where the true treatment effect is likely to lie. Bayesian methods are particularly useful when there is prior knowledge about the treatment effect or when quantifying the uncertainty of the estimates is crucial.

The latest opportunities are offered by the application of Large Language Models (LLMs) in this context. LLMs offer significant potential to enhance experimental designs aimed at exploring heterogeneity in treatment effects. By leveraging LLMs, researchers can generate synthetic participants to complement real study participants, addressing issues of underrepresentation and allowing for the exploration of rare trait combinations. This capability enables better covariate balancing and more robust model validation. LLMs can be integrated into adaptive randomization strategies, propensity score modeling, and outcome prediction, creating a multi-stage process that dynamically improves as the study progresses. These models can generate hypothetical scenarios, identify confounding factors, and refine propensity score models, ultimately improving the allocation of participants to different interventions. Furthermore, LLMs can simulate potential outcomes, aiding in sequential randomization and response-adaptive allocation. This approach allows for a more flexible and efficient exploration of treatment effect heterogeneity, as the experimental design can be continuously updated based on both real and synthetic data insights. The integration of LLMs in this context promises to enhance study efficiency and the validity of findings regarding heterogeneous treatment effects, potentially revolutionizing how we design and conduct experiments.

Heterogeneity, mechanisms and personalisation

Uncovering heterogeneity in the uptake of behavioral interventions helps us comprehend the varying effectiveness of these interventions, which sheds light on different operational

mechanisms of BIs. This understanding not only clarifies how these tools function, leading to conceptual refinement, but also enables us to customize and personalize these interventions. For example, Krefeld-Schwalb et al. (2024) through a series of large-scale online as well as offline experiments, underscore the need to understand omitted moderators which can explain why treatments might vary in their implementation intensity. Understanding this segmentation can, in turn, enable practitioners to choose between policies.

Personalisation methods vary but can be broadly classified, as we propose here, into top-down and bottom-up ways. A top-down approach to personalisation relies on utilizing behavioral profiling. In this approach, clusters of individuals are first identified using previously available information on behavioral characteristics, such as one's demographics, socio-economic preferences, cognitive abilities and so on. Using different clustering algorithms (Nikoloski et al., 2024), it is possible to uniquely identify different clusters and thereby predict underlying economic and cognitive barriers that hinders the uptake of desirable behaviors, thereby assigning specific behavioral interventions to these different clusters. The top-down approach to personalisation is often synonymous to ex-ante personalisation, that is personalisation before the delivery of an intervention. Contrary to this, a bottom-up approach relies on utilizing response efficacy. In this approach, a generic intervention is first administered to all individuals. Following this, average causal effects of the treatment are measured across different clusters or groups of individuals using computational approaches such as heterogeneity analysis via causal forests (Veltri, 2023). Identification of heterogeneous treatment effects enables us to uniquely determine what works best and for whom and tailor behavioral interventions based on such a ranking. This bottom-up data driven approach is often synonymous to ex-post personalisation, that is personalisation after the delivery of the intervention.

By employing computational social science techniques, we can ultimately create meta-rules that enable practitioners to classify and design alternative behavioral toolkits in a streamlined and practical manner (Banerjee and Galizzi, 2024). This involves accurately identifying subpopulations with varying treatment effects and uncovering previously unrecognized sources of heterogeneity. For example, following field experiments on using reminders to improve the uptake of student financial aid, Athey et al. (2023b) applied a bottom-up approach finding that text and email reminders worked best for students who were already somewhat predisposed to applying for financial aid. In contrast, students who were less likely to file for aid remained largely unaffected by these reminders. Based on this, they suggest avoiding expensive efforts to engage individuals who are unlikely to respond. Similarly, in the context of modern contraceptive methods, Athey et al. (2023a) suggest that "low-cost individualized recommendations can potentially be as effective in increasing unfamiliar technology adoption as providing large subsidies." While the evidence on the benefits of computational social science methods in behavioral science is growing, direct tests of personalized interventions vs. "one-size-fits-all" policies are largely missing.

Implementing computational methods for personalizing behavioral interventions raises ethical concerns. While personalisation can improve effectiveness, it risks reinforcing societal inequalities if certain groups are excluded based on predicted response rates. This "optimization-fairness trade-off" could neglect vulnerable populations. Moreover, there are questions about algorithmic transparency and accountability, as practitioners and subjects must understand intervention assignments. Additionally, extensive data collection may infringe on privacy rights and autonomy, reinforcing claims of a "nanny-state" government. Addressing these challenges requires clear governance frameworks that balance optimisation with equity, such as equity audits of algorithms and transparent processes for individuals to understand and contest their intervention assignments.

Overall, we advocate for a greater utilization of these machine learning methods to harness the diversity within behavioral public policy.

Author contributions

SB: Conceptualization, Project administration, Writing – original draft, Writing – review & editing. GV: Conceptualization, Project administration, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The authors thank Adam Oliver, Eyal Pe'er, Matteo M. Galizzi, Peter John for comments on an earlier draft of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Athey, S., Bergstrom, K., Hadad, V., Jamison, J. C., Özler, B., Parisotto, L., et al. (2023a). Can personalized digital counseling improve consumer search for modern contraceptive methods? *Sci. Adv.* 9:eadg4420. doi: 10.1126/sciadv.adg4420

Athey, S., and Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci.* 113, 7353–7360. doi: 10.1073/pnas.1510489113

Athey, S., Keleher, N., and Spiess, J. (2023b). Machine learning who to nudge: causal vs predictive targeting in a field experiment on student financial aid renewal. *arXiv* [Preprint]. arXiv:2310.08672. doi: 10.48550/arXiv.2310.08672

Banerjee, S., and Galizzi, M. M. (2024). "People are different! And so should be behavioural interventions," in *The Behavioral Economics Guide 2024*, ed. A. Samson, 109–118.

Banerjee, S., Galizzi, M. M., John, P., and Mourato, S. (2023). Sustainable dietary choices improved by reflection before a nudge in an online experiment. *Nat. Sustain.* 6, 1632–1642. doi: 10.1038/s41893-023-01235-0

Banerjee, S., Grüne-Yanoff, T., John, P., and Moseley, A. (2024). It's time we put agency into behavioural public policy. *Behav. Public Policy* 8, 789–806. doi: 10.1017/bpp.2024.6

Banerjee, S., and John, P. (2024). Nudge plus: incorporating reflection into behavioral public policy. *Behav. Public Policy* 8, 69–84. doi: 10.1017/bpp.2021.6

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2020). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Bayesian Anal.* 15, 965–1056. doi: 10.1214/19-BA1195

Hertwig, R. (2017). When to consider boosting: some rules for policy-makers. *Behav. Public Policy* 1, 143–161. doi: 10.1017/bpp.2016.14

Hertwig, R., and Grüne-Yanoff, T. (2017). Nudging and boosting: steering or empowering good decisions. *Perspect. Psychol. Sci.* 12, 973–986. doi: 10.1177/1745691617702496

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *J. Comput. Graph. Stat.* 20, 217–240. doi: 10.1198/jcgs.2010.08162

John, P., Smith, G., and Stoker, G. (2009). Nudge nudge, think think: two strategies for changing civic behaviour. *Polit. Q.* 80, 361–370. doi: 10.1111/j.1467-923X.2009.02001.x

Krefeld-Schwalb, A., Sugerman, E. R., and Johnson, E. J. (2024). Exposing omitted moderators: explaining why effect sizes differ in the social sciences. *Proc. Natl. Acad. Sci.* 121:e2306281121. doi: 10.1073/pnas.2306281121

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Natl. Acad. Sci.* 116, 4156–4165. doi: 10.1073/pnas.1804597116

Nie, X., and Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika* 108, 299–319. doi: 10.1093/biomet/asaa076

Nikoloski, M., Botzen, W. J. W., Talevi, M., Blasch, J., Poblete-Cazenave, M., Banerjee, S., et al. (2024). Methods to Optimise and Personalise Residential Energy Efficiency Interventions: *Review of the Literature*. Working Paper Version.

Thaler, R. H., and Sunstein, C. R. (2008). Nudge: Improving *Decisions about Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.

Veltri, G. A. (2023). Harnessing heterogeneity in behavioural research using computational social science. *Behav. Public Policy* 1–18. doi: 10.1017/bpp.2023.35. [Epub ahead of print].