Prediction, Projection and Performativity

Konstantin Genin, Sebastian Zezulka

Epistemology and Ethics of Machine Learning
University of Tübingen

ethics.epistemology.ai

machine learning
new perspectives
for science

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Research Group: Epistemology and Ethics of Machine Learning

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Konstantin Genin
Group Leader

Inherent complexity of scientific problems. Interactions between morals and methodology.

Raysa Benatti
PhD student

AI policy evaluation. Algorithmic prediction and gender-based violence.

Mykhailo Bogachov
Visiting PhD. Student

Normative implications of performativity in AI systems. Effects of LLMs on moral reasoning.

Sebastian Zezulka
PhD. Student

Dynamical perspectives on algorithmic justice.

Prediction-Allocation Problems

Setting: policy problems in which social-statistical **predictions** are used to determine the **allocation** of social goods.

Interested in how methodologies of prediction structure

(1) distributive **outcomes** and

(2) **discourses** about which distributive outcomes are possible and desirable.

Risk Assessment and Public Employment Services

The algorithm takes as **input**

- the education and employment history (X),
- and gender (A)

of a recently unemployed person, and **outputs** a prediction ($\hat{\mathrm{Y}}$) of (the risk of) long-term unemployment (Y).

On the basis of the prediction ($\hat{\mathrm{Y}}$), a case-worker assigns the person to some labor-market program (D) that is causally relevant for their employment prospects (Y) .

Sebastian Zezulka and Konstantin Genin (2024). "From the fair distribution of predictions to the fair distribution of social goods." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.

Risk Assessment and Public Employment Services

The risk score may support a number of different policies.

- In Belgium: individuals at high risk of long-term unemployment are **prioritized** (Desiere and Struyven, 2020).

- In Austria: risk scores classify the recent unemployed into those with (i) good prospects in the next six months; (ii) bad prospects in the next two years; and (iii) everyone else. Support measures **target** the third group, while offering **only limited support** to the first and second group (Allhutter et al., 2020).

Allhutter, D., Cech, F., Fischer, F., Grill G, and Mager, A. (2020) Algorithmic profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective.

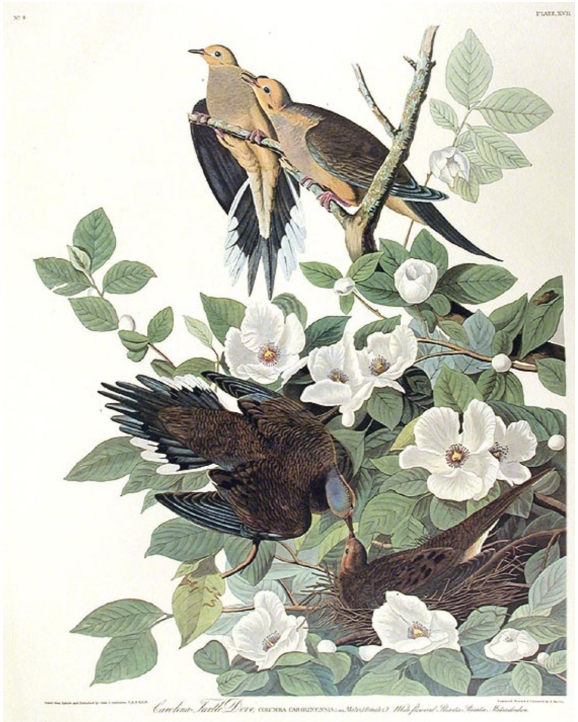
Sam Desiere and Ludo Struyven. Using Artificial Intelligence to classify Jobseekers: The Accuracy-Equity Trade-off. Journal of Social Policy, 50(2):367–385, 2020.

Hawks and Doves

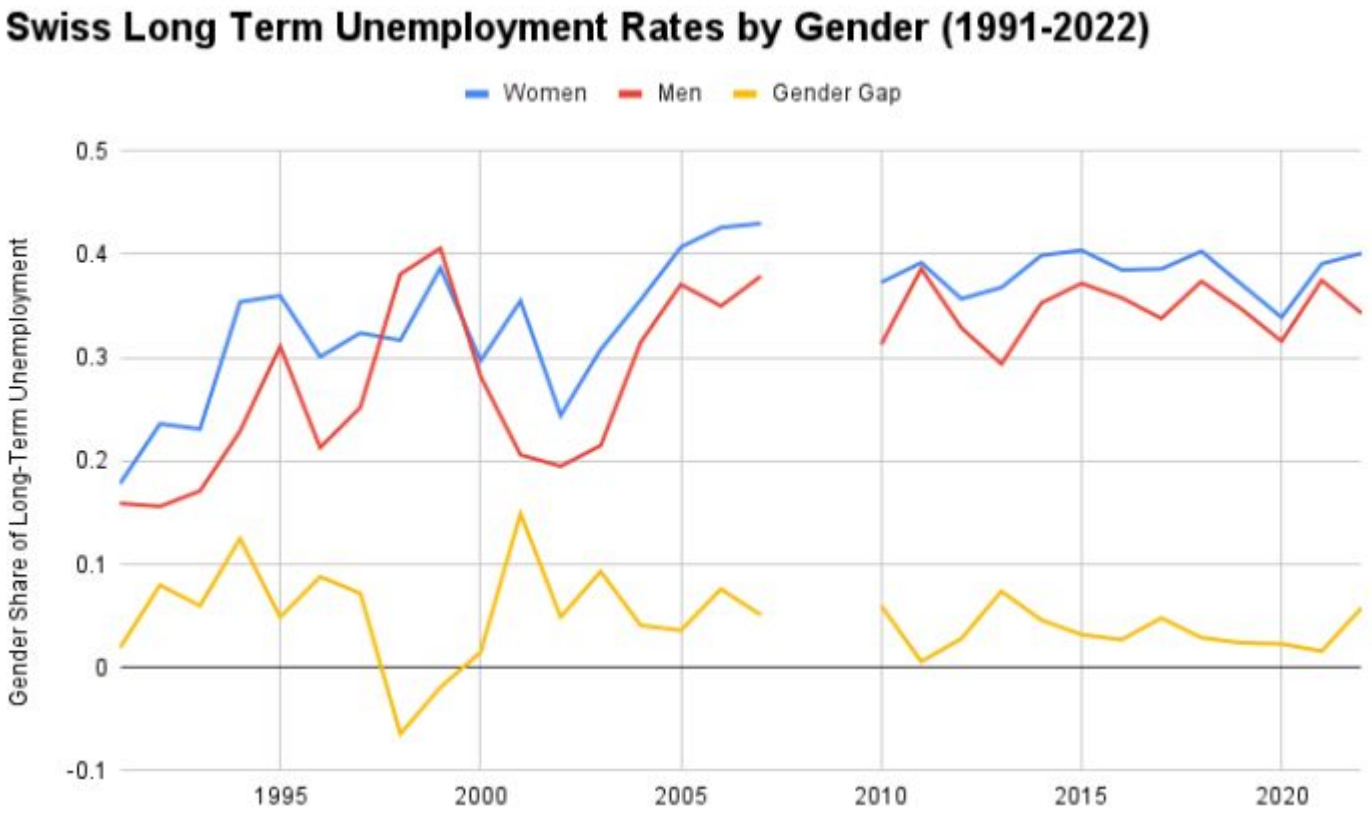
Advocates of the Austrian policy argue in terms of *efficiency.*

Hawks and Doves

Critics of the Austrian plan worry about exacerbating long-standing structural inequalities in the labor market, preferring to focus assistance on the worst-off.



The Gender Reemployment Gap: Switzerland

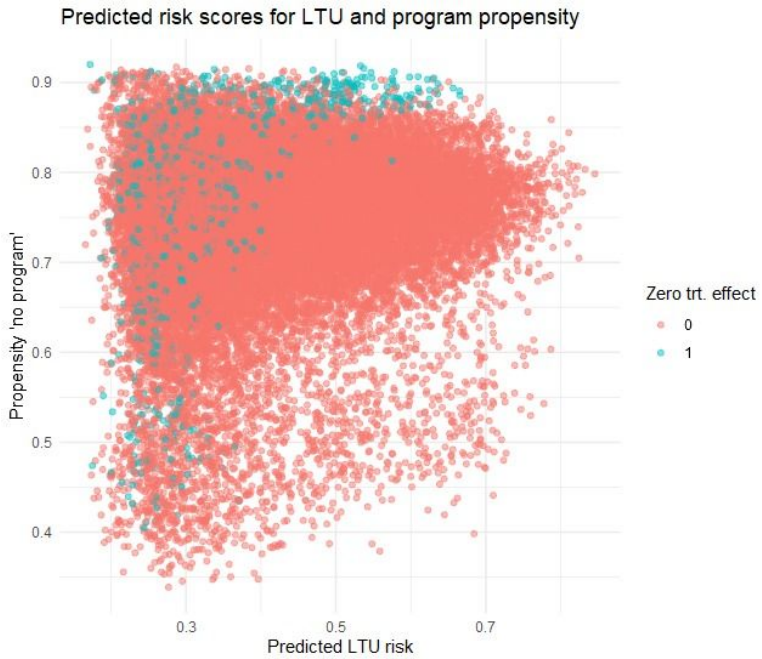


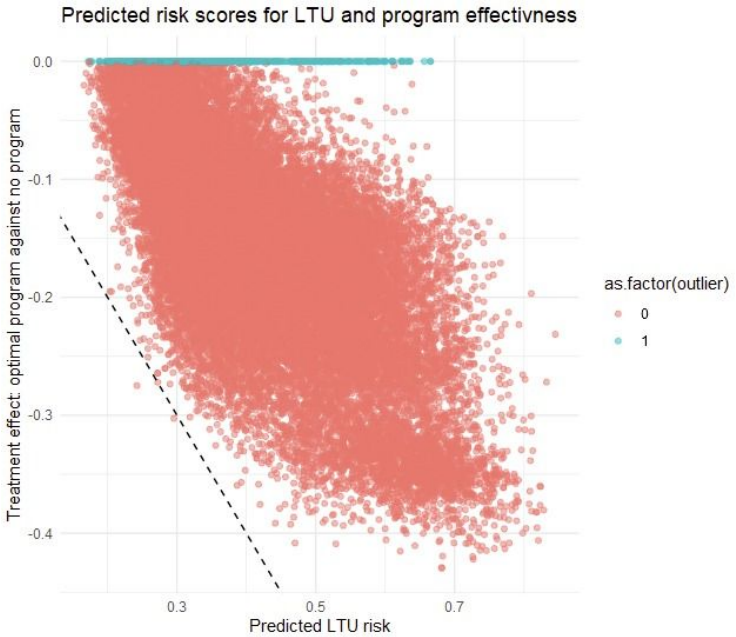
Risk Assessment and Public Employment Services

What if the highest risk are those simply more likely, historically, to get the least effective programs?

Risk Assessment and Public Employment Services

What if the highest risk are those simply more likely, historically, to get the least effective programs?





Risk Assessment and Public Employment Services

What if the highest risk are those simply more likely, historically, to get the least effective programs?

A case of **self-fulfilling** prophecy: goods are withheld from individuals who are considered high-risk (because historically they would have been denied these goods) and whose subsequent poor outcomes are due, in part, to their smaller share of goods.

Risk Assessment and Public Employment Services

What if the highest risk are those simply more likely, historically, to get the least effective programs?

A case of **self-fulfilling** prophecy: goods are withheld from individuals who are considered high-risk (because historically they would have been denied these goods) and whose subsequent poor outcomes are due, in part, to their smaller share of goods.

Note that this could happen even when the predictions are *accurate.*

Pittsburgh Hospital Admissions

The algorithm takes as **input**

- medical history and vital signs (X)

of an individual presenting at the hospital with pneumonia, and **outputs** a prediction $(\hat{Y})$ of (the risk of) mortality (Y).

On the basis of the prediction $(\hat{Y})$ physicians make a decision (D) whether to hospitalize or follow-up as outpatient. The decision (D) is causally relevant to their survival (Y) .

Cooper, Gregory F., et al. (1997) "An evaluation of machine-learning methods for predicting pneumonia mortality." *Artificial intelligence in medicine* 9.2: 107-138.

Tal, Eran. (2023) "Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Pittsburgh Hospital Admissions

The algorithm learns that having asthma *lowers* the risk of mortality of pneumonia patients.

It's true! Such patients were often admitted directly into the intensive care unit, where they received aggressive care. This reduced the mortality rate of asthmatics with pneumonia relative to the overall pneumonia patient population.

Tal, Eran. (2023) "Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Pittsburgh Hospital Admissions

The algorithm learns that having asthma *lowers* the risk of mortality of pneumonia patients.

But if risk scores were naively used for triage, the effects could have been **disastrous**.

Tal, Eran. (2023) "Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Pittsburgh Hospital Admissions

The algorithm learns that having asthma *lowers* the risk of mortality of pneumonia patients.

A case of **self-defeating prophecy**: a good is denied to those considered low risk (because historically they would have received this good) and whose subsequent poor outcomes are due, in part, to their smaller share of the good.

Tal, Eran. (2023) "Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Pittsburgh Hospital Admissions

The algorithm learns that having asthma *lowers* the risk of mortality of pneumonia patients.

Note that this could happen even though the predictions are *accurate* and **everyone is behaving fairly and responsibly.**

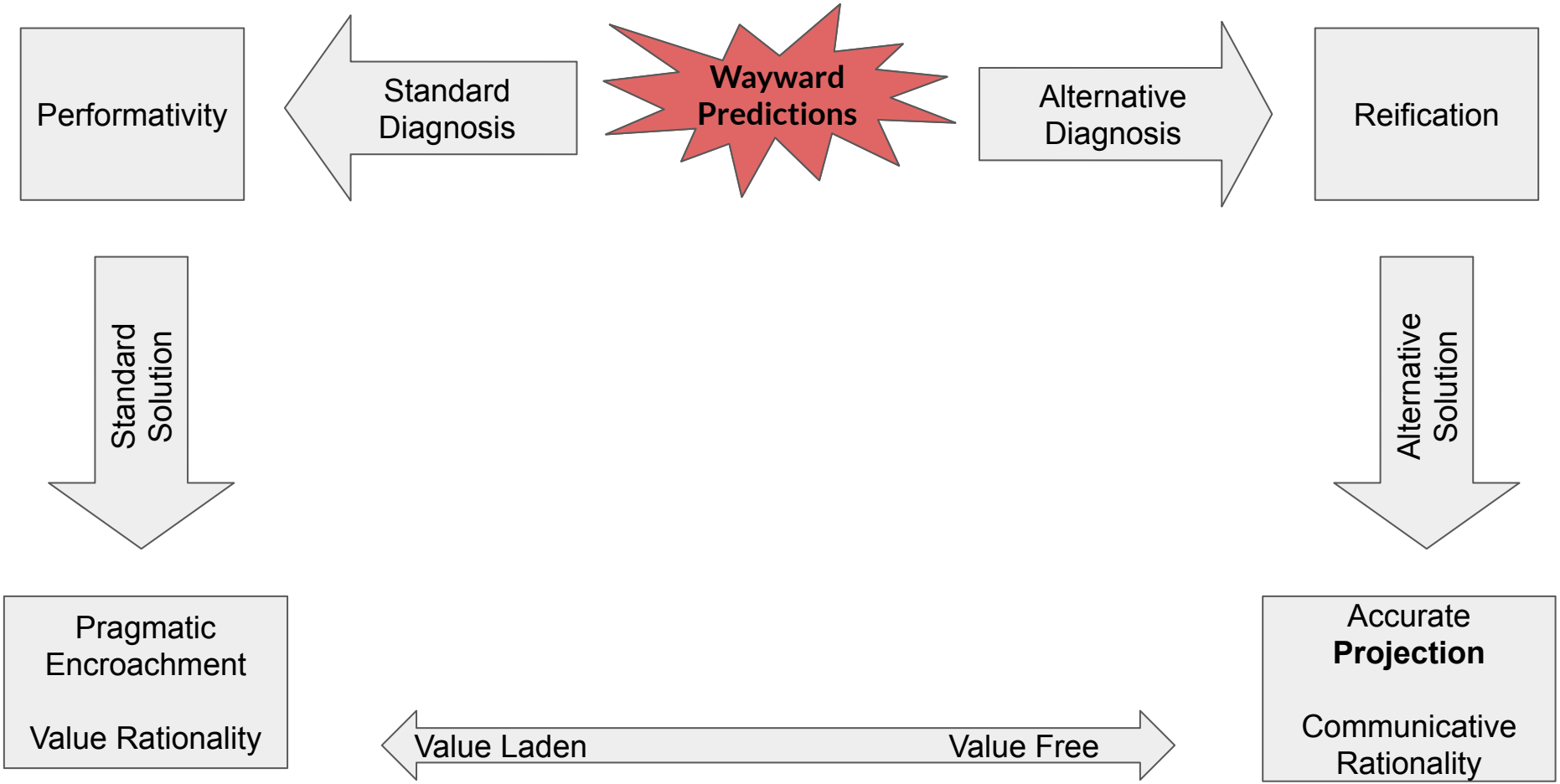
Tal, Eran. (2023) "Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*.

Prediction-Allocation Problems

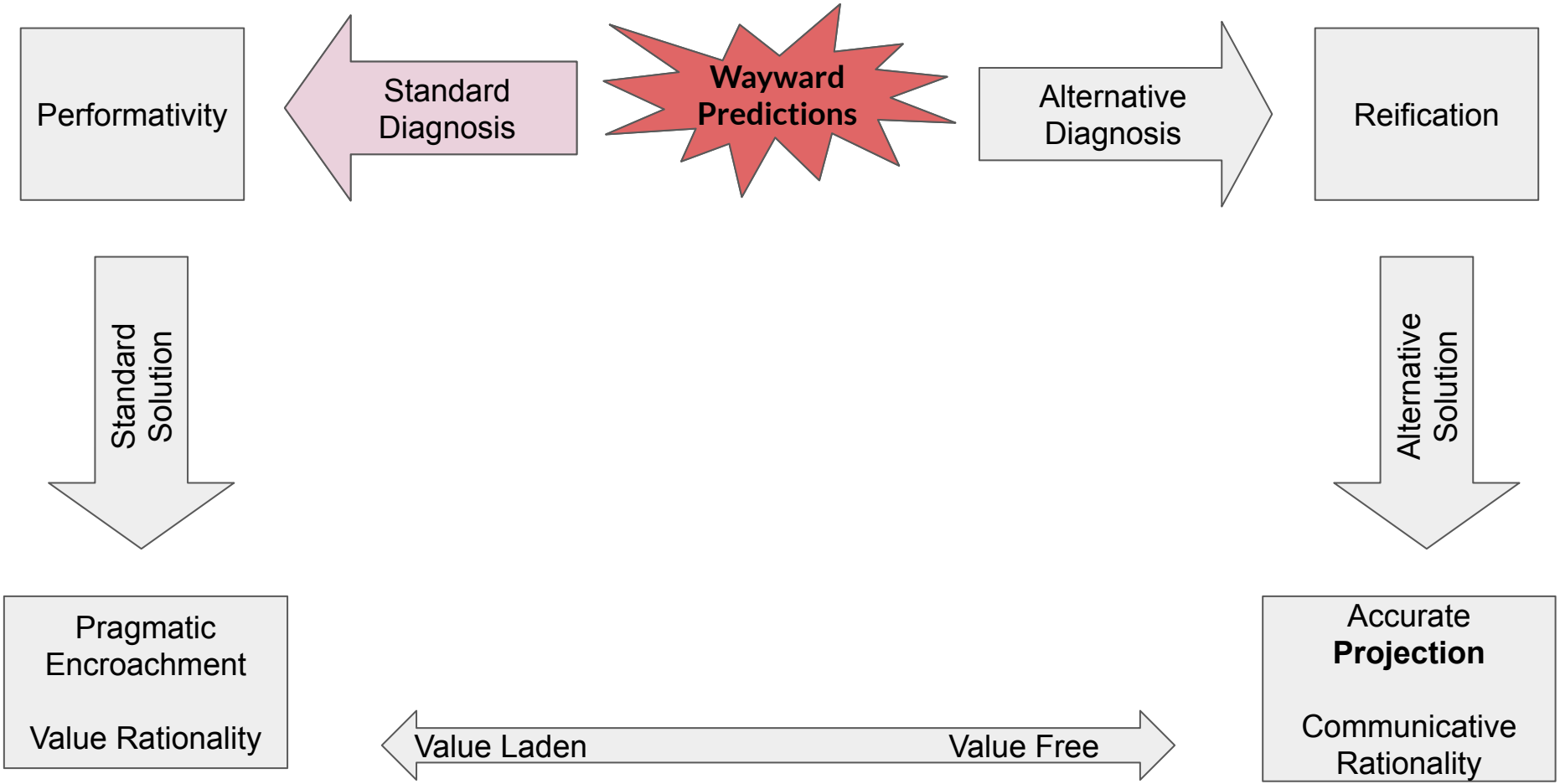
What is going wrong in these prediction-allocation settings?

Why are problems arising despite accurate social-scientific predictions?

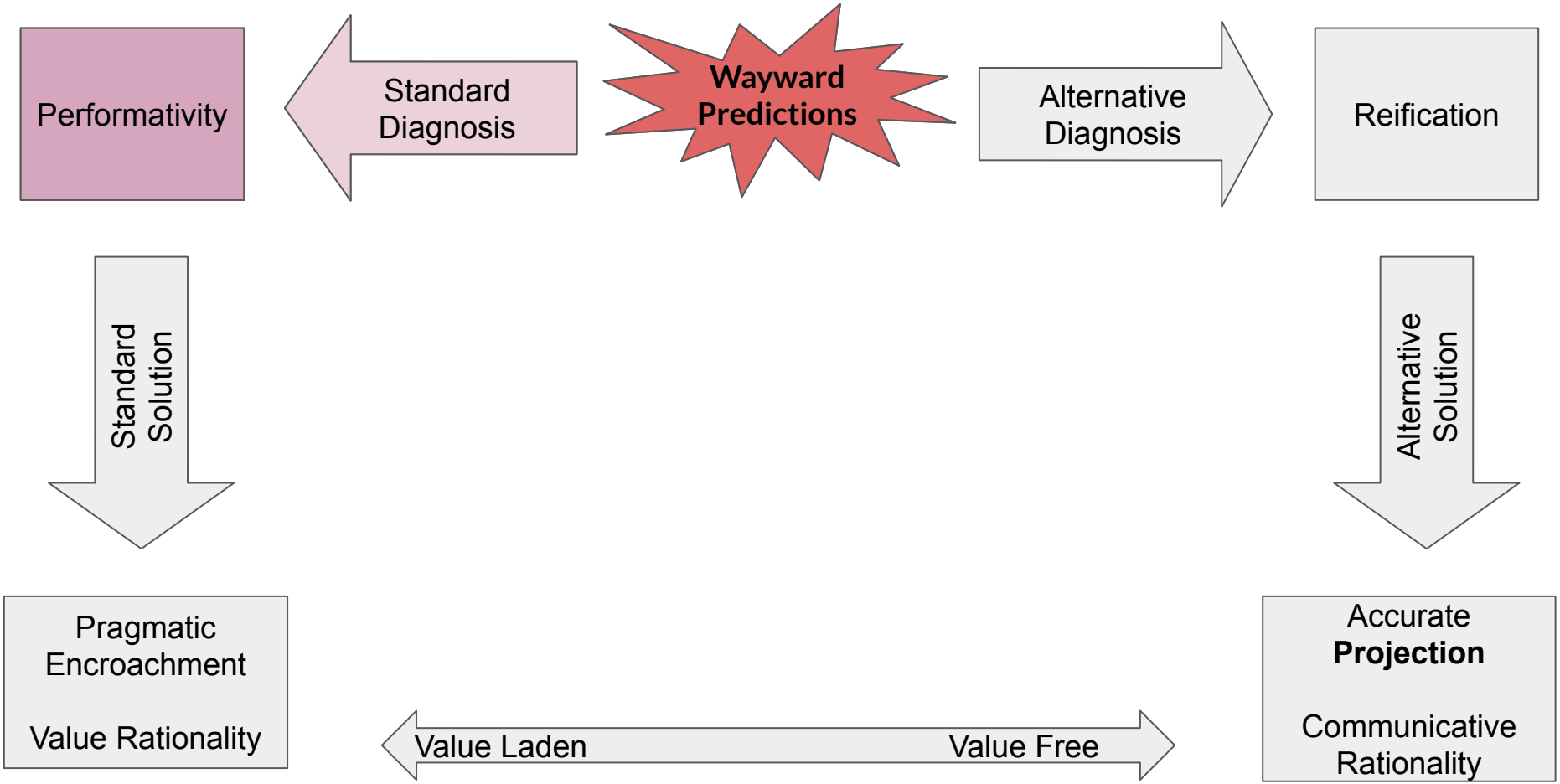
Wayward Predictions



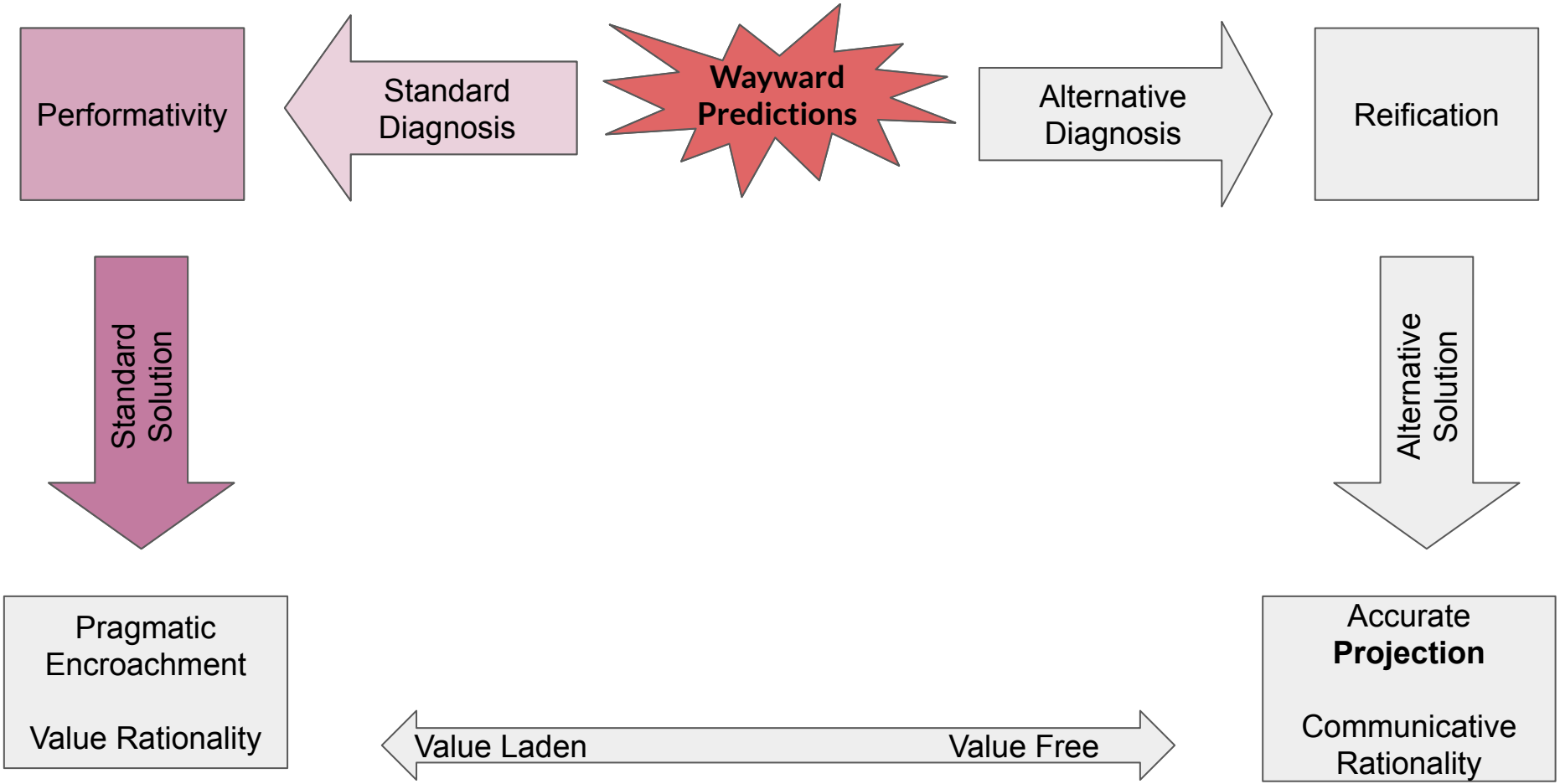
Wayward Predictions



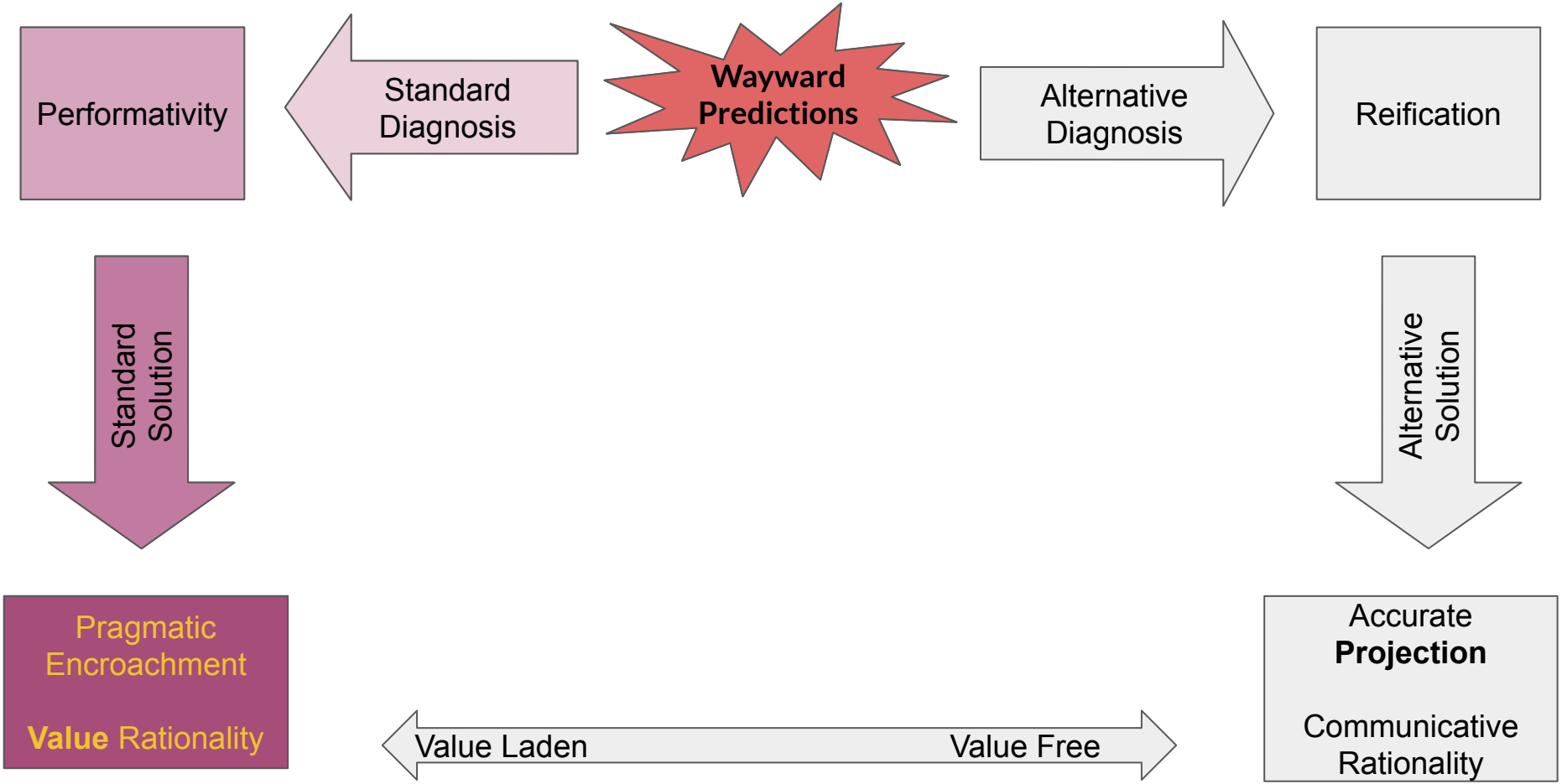
Wayward Predictions



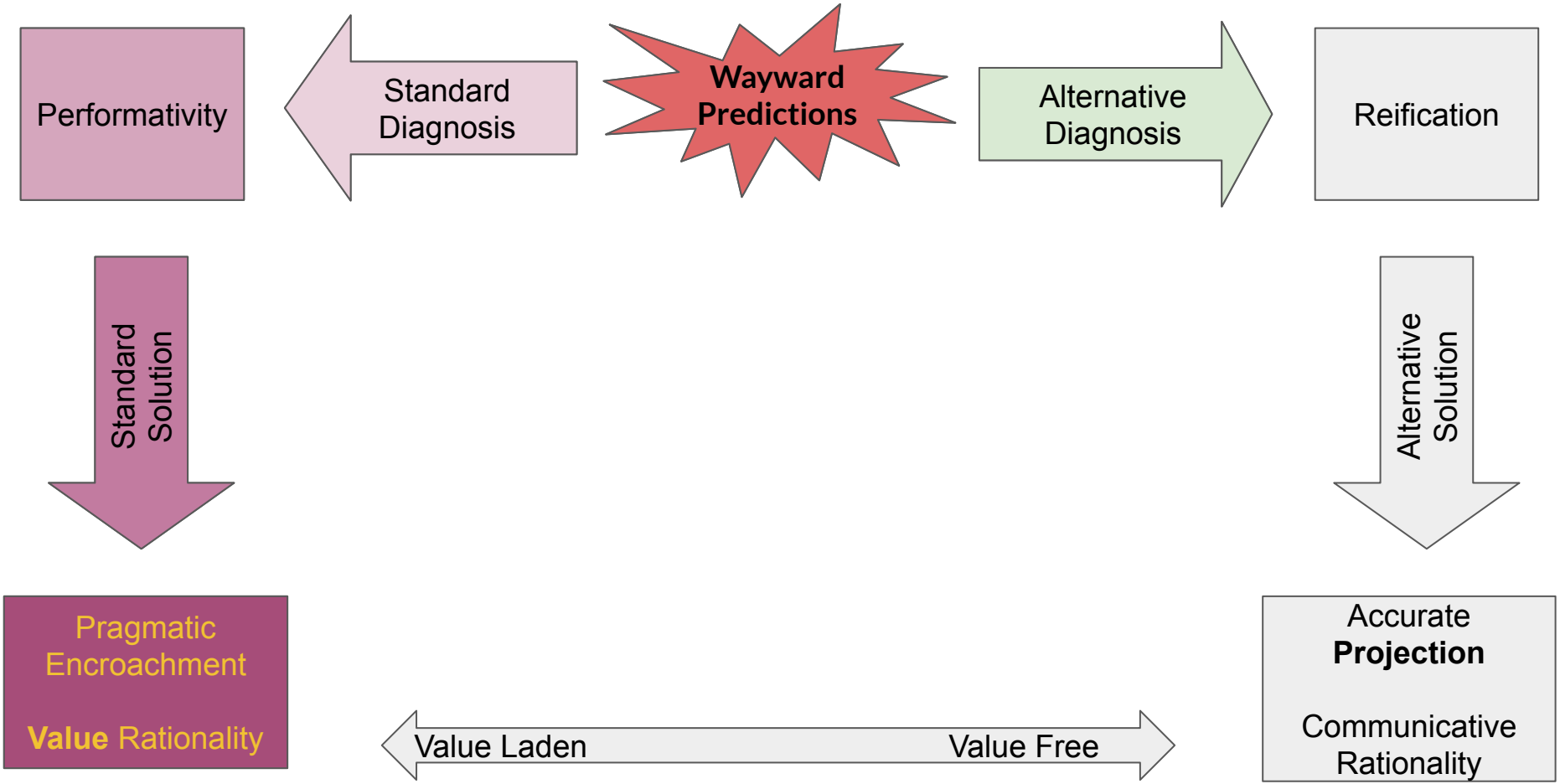
Wayward Predictions



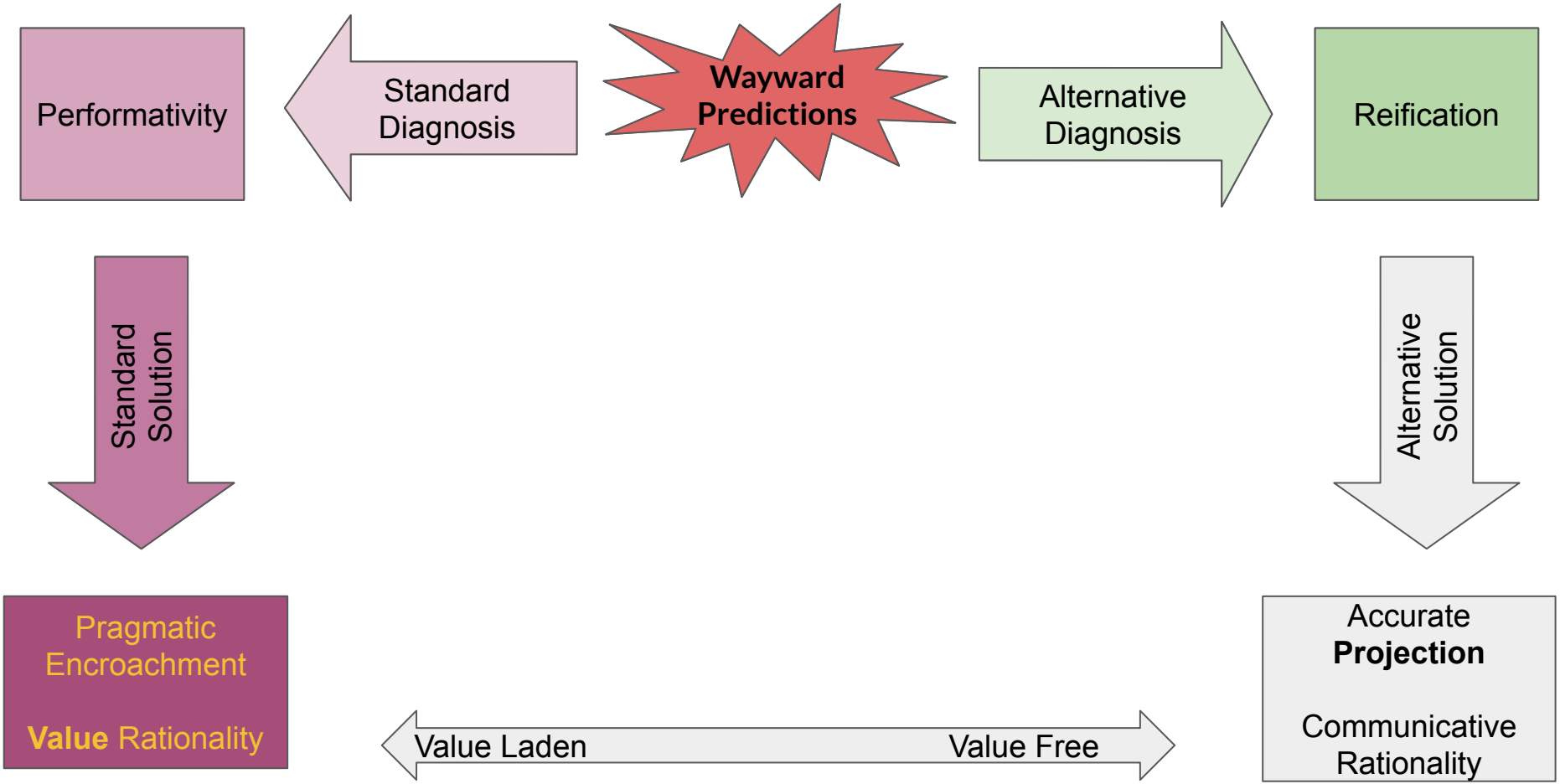
Wayward Predictions



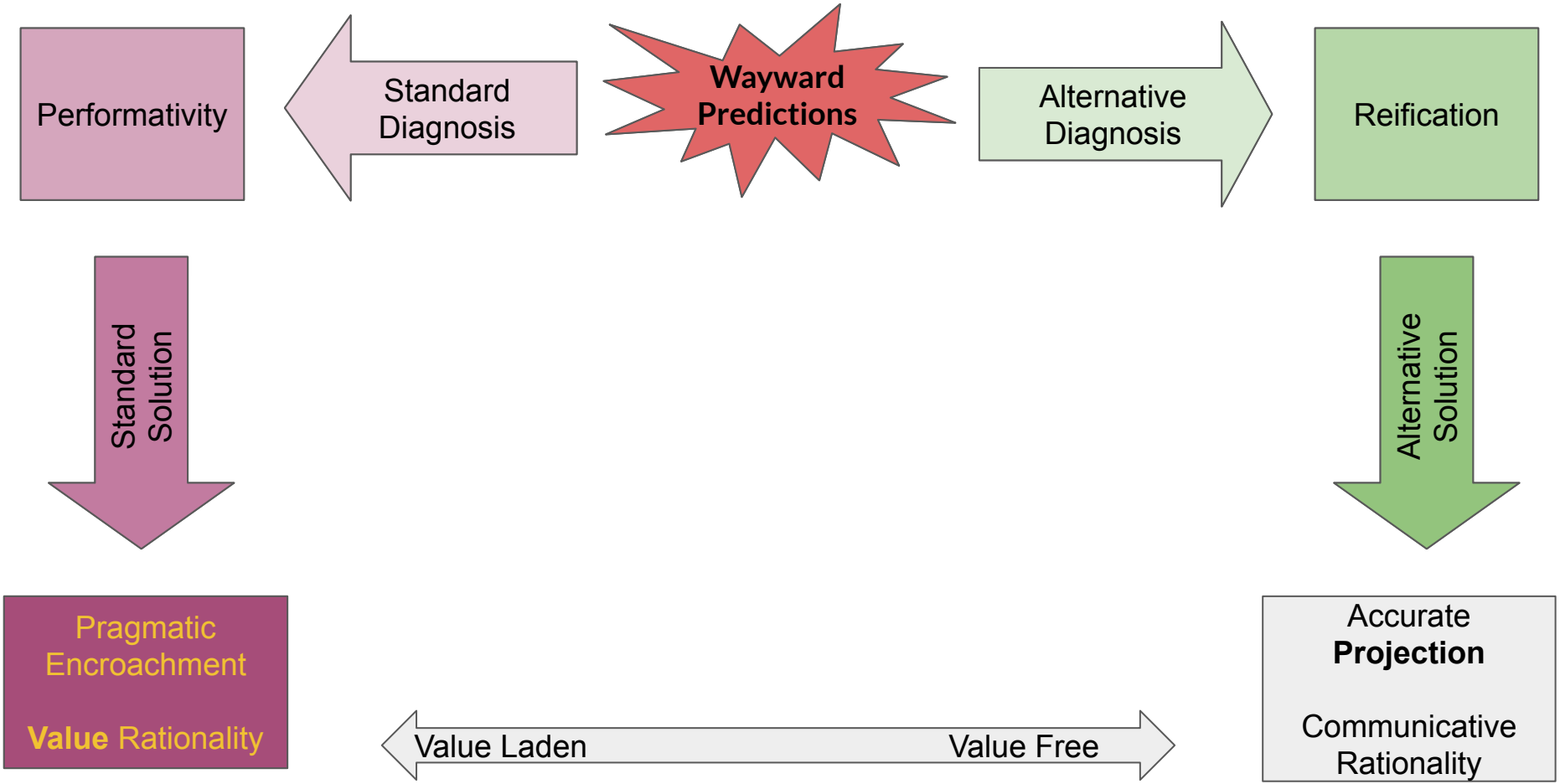
Wayward Predictions



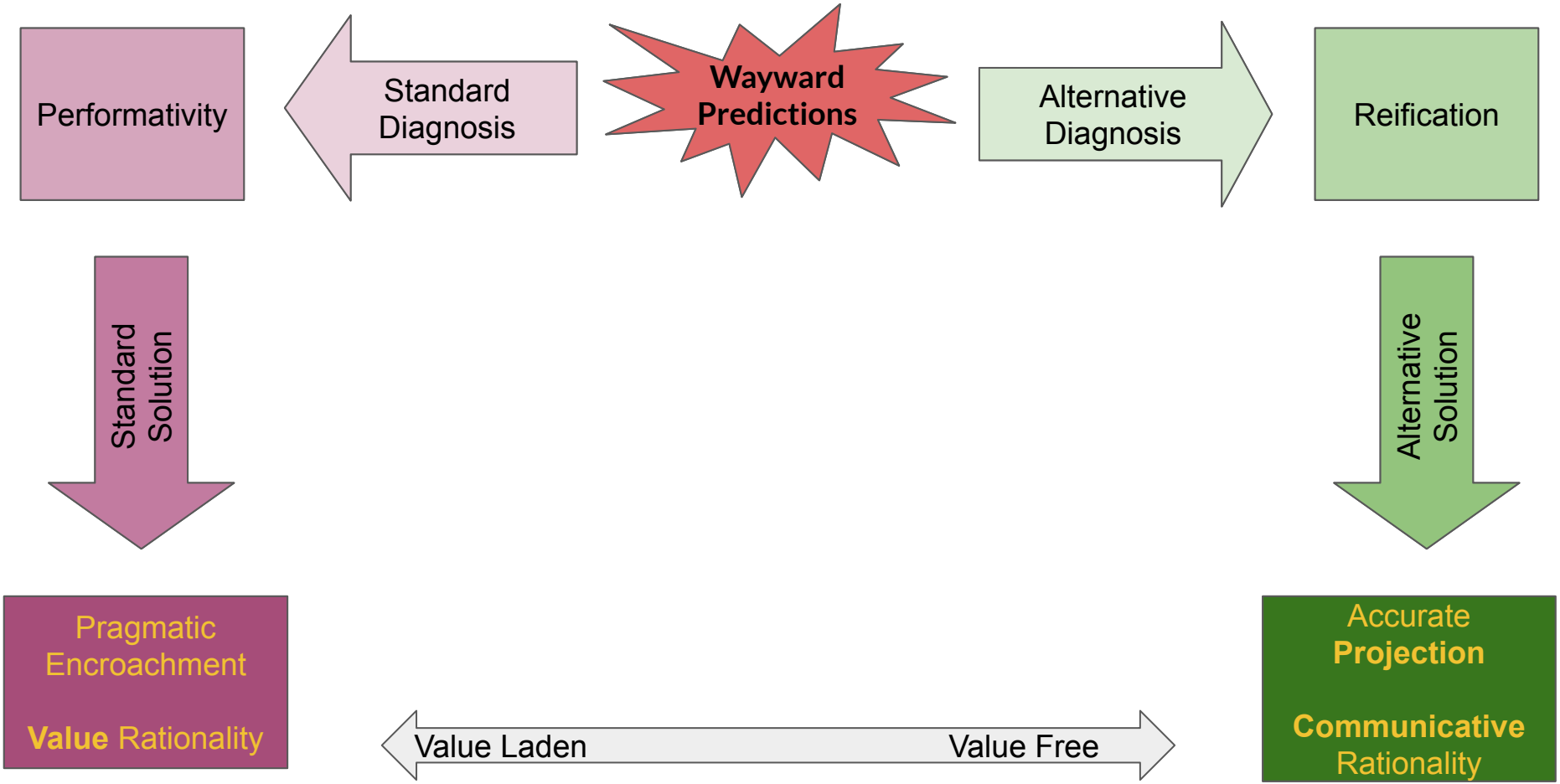
Wayward Predictions



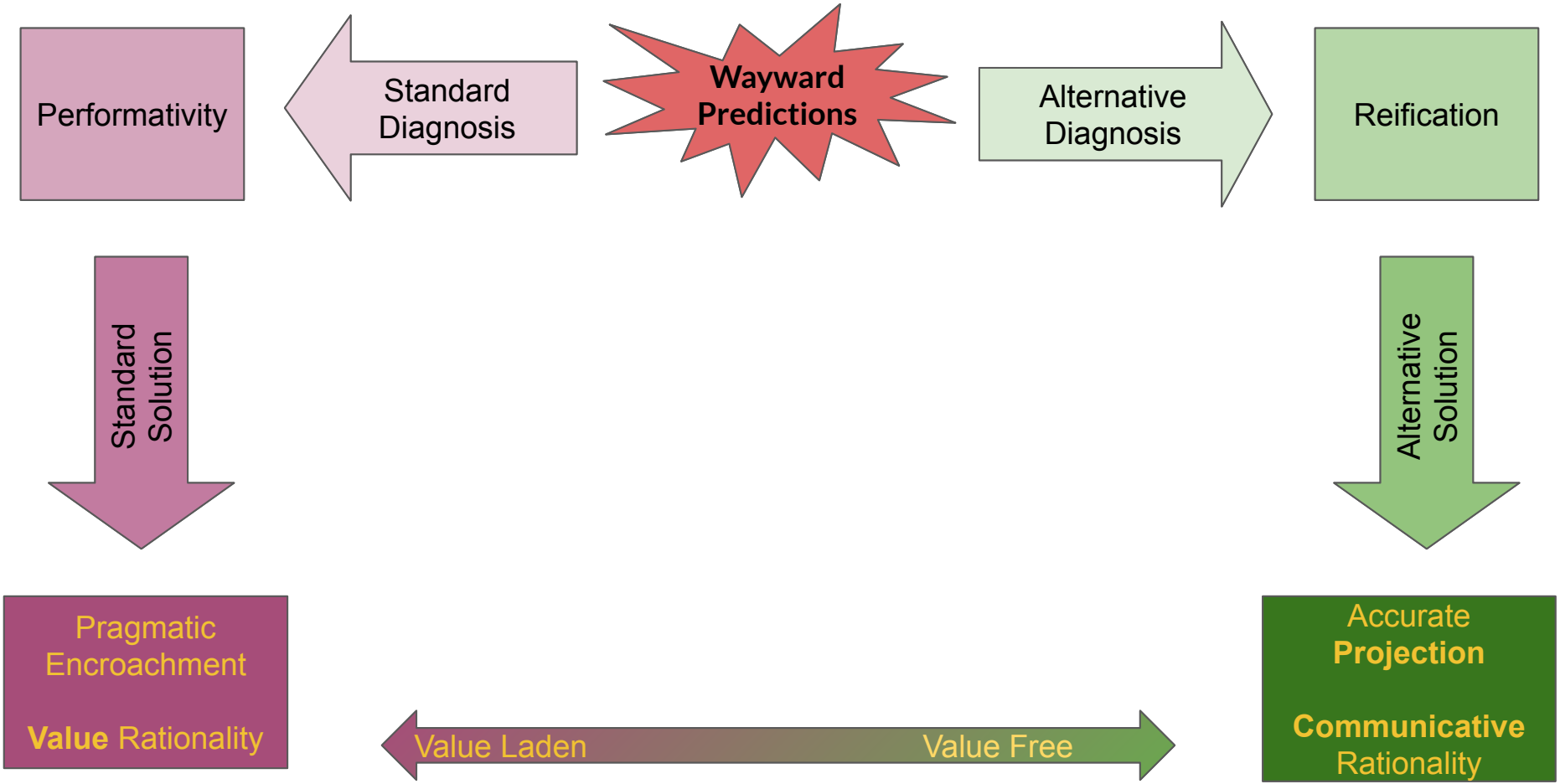
Wayward Predictions



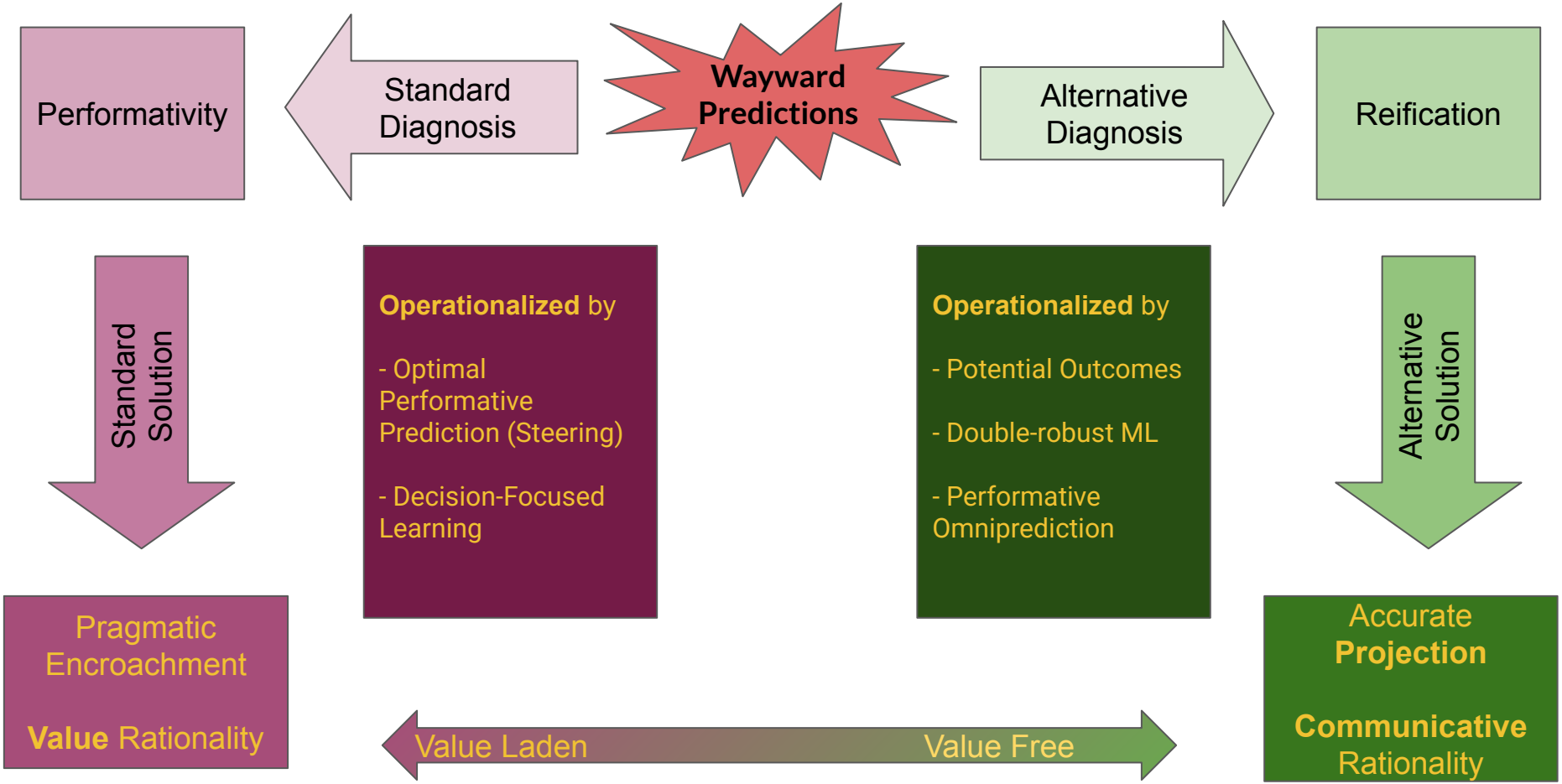
Wayward Predictions



Wayward Predictions



Wayward Predictions



Performativity: Macro

A model is ***performative*** when, in addition to serving epistemic purposes, it causally influence the system that it is meant to represent.

Performativity is ubiquitous in social science.

Individuals may react to the creation of a social kind (e.g. multiple-personality disorder) by identifying with the new classification, thereby changing their behavior (Hacking, 1996).

Hacking, Ian (1995). “The looping effects of human kinds.” In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 351–394).

Performativity: Macro

A model is ***performative*** when, in addition to serving epistemic purposes, it causally influence the system that it is meant to represent.

Performativity is ubiquitous in social science.

Individuals may react to predictions made by models of viral spread by limiting social interaction (Van Basshuysen et al., 2021).

Van Basshuysen et al. (2021). “Three Ways in which Pandemic Models May Perform a Pandemic.” *Erasmus Journal for Philosophy and Economics* 14 (1): 10-127.

Performativity: Macro

A model is ***performative*** when, in addition to serving epistemic purposes, it causally influence the system that it is meant to represent.

Performativity is ubiquitous in social science.

Individuals may react to a new credit scoring system by strategically manipulating the way they are represented in data (Hu et al., 2021).

Hu, Lily, Nicole Immorlica, and Jennifer Wortman Vaughan (2019). "The disparate effects of strategic manipulation." *Proceedings of the Conference on Fairness, Accountability, and Transparency*.

Performativity: Macro and Micro

These are instances of **macro**-performativity, in which models have diffuse effects on entire populations. These effects are often unforeseen, unintended and difficult to anticipate.

We are interested in **micro**-performativity, in which a local prediction about an individual causally influences that individuals outcomes. These effects are typically **intended** and (somewhat) easier to anticipate (Zezulka and Genin, 2023; Kim and Perdomo, 2022).

Kim, Michael P., and Juan C. Perdomo. (2022) "Making decisions under outcome performativity."

Zezulka, S. and Genin, K. (2023). “Performativity and Prospective Fairness” *Fairness Through the Lens of Time*, NeurIPS 2023.

Performativity and Wayward Prediction

Performativity clearly has something to do with cases of wayward prediction.

- Predictions of low mortality (for asthmatics) are perversely causing them to have worse outcomes.

- Predictions of long-term unemployment (e.g., for women) are perversely causing them to have worse employment outcomes.

An intuitive suggestion: dealing with wayward prediction is a matter of **managing** performativity.

Pragmatic Encroachment

In epistemology, the pragmatic ‘encroaches’ on the epistemic when practical considerations, such as the severity of the consequences foreseeable errors, make a difference to whether some agent **knows** a proposition (Stanley, 2005).

Stanley, Jason (2005). Knowledge and practical interests. *Oxford University Press*.

Pragmatic Encroachment

In epistemology, the pragmatic ‘encroaches’ on the epistemic when practical considerations, such as the severity of the consequences of foreseeable errors, make a difference to whether some agent **knows** a proposition (Stanley, 2005).

We collect a number of proposals for managing performativity under the heading of **pragmatic encroachment:** they all suggest that pragmatic considerations, and not just epistemic considerations of ‘accuracy’, should inform predictions in performative settings.

Stanley, Jason (2005). Knowledge and practical interests. *Oxford University Press*.

Inductive Risk (in Performative Contexts)

“The scientist acting as advisor should consider … the possible consequences of **incorrectly** accepting or rejecting the claim, and they should weigh the importance of the uncertainties accordingly” (2008, p. 81).

Douglas, Heather (2008). Science, Policy and the Value-Free Ideal. *University of Pittsburgh Press*.

Inductive Risk (in Performative Contexts)

“The scientist acting as advisor should consider … the possible consequences of **incorrectly** accepting or rejecting the claim, and they should weigh the importance of the uncertainties accordingly” (2008, p. 81).

But if the policy context is unclear, this advice is very difficult to act on!

Douglas, Heather (2008). Science, Policy and the Value-Free Ideal. *University of Pittsburgh Press*.

Inductive Risk (in Performative Contexts)

Suppose we are

- developing an algorithm to predict long-term unemployment
- and worried about exacerbating the gender reemployment gap.

Simultaneously, a debate is raging is policy circles. Should we implement

- an Austrian-type triage scheme, focusing on *efficiency;*
- or a Flanders-type prioritarian scheme, focusing on demands of *justice*.

Should we worry more about **under-** or **over-**estimating the risk of long-term unemployment for women?

Inductive Risk (in Performative Contexts)

This is essentially a version of Jeffrey's response to Rudner:

"It is certainly meaningless to speak of *the* cost of mistaken acceptance or rejection, for by its nature a putative scientific law will be relevant in a great diversity of choice situations among which the cost of a mistake will vary greatly" (1956, 242).

Jeffrey, Richard C. (1956) "Valuation and Acceptance of Scientific Hypotheses." *Philosophy of Science*. 23(3):237-246.

Inductive Risk (in Performative Contexts)

“The scientist should not think about the potential consequences of making an **accurate** empirical claim and slant their advice accordingly” (2008, p. 81; emphasis mine).

Douglas, Heather (2008). Science, Policy and the Value-Free Ideal. *University of Pittsburgh Press*.

Inductive Risk (in Performative Contexts)

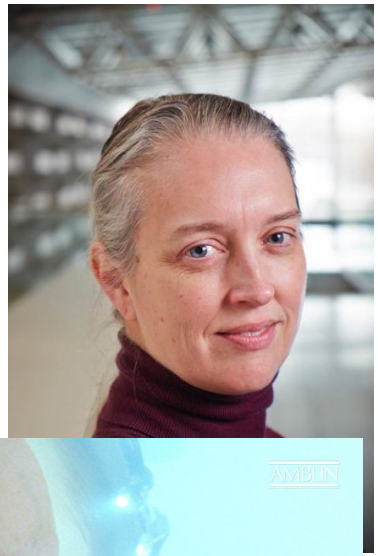
“The scientist should not think about the potential consequences of making an **accurate** empirical claim and slant their advice accordingly” (2008, p. 81; emphasis mine).

But in cases of self-fulfilling prophecy we are not worried about being in **error** but rather being **correct for the wrong reason!**

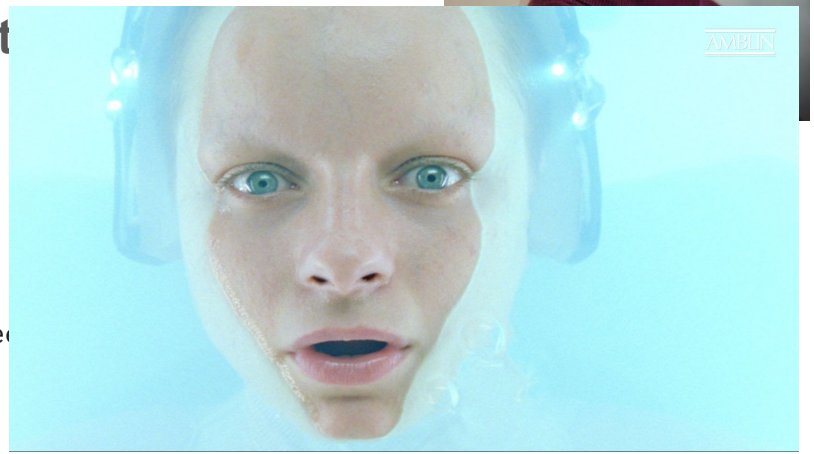
Douglas, Heather (2008). Science, Policy and the Value-Free Ideal. *University of Pittsburgh Press*.

Inductive Risk (in Performative Contexts)

“The scientist should not think about the potential consequences of making an **accurate** empirical claim and slant their advice accordingly” (2008, p. 81; emphasis mine).



But in cases of self-fulfilling prophecy we are not worried about being in **error** but rather being **correct** for the wrong **reason!**



Douglas, Heather (2008). Science, Policy and the Value-Free

Inductive Risk (in Performative Contexts)

“... the book focuses exclusively on the natural science as a source of desired expertise. My neglect of the social sciences, such as psychology, sociology, and economics, arises partly from … unique problems of reflexivity, as the subjects of the research can read and understand the research, and alter their behavior as a result. How the ideas I develop here would apply to such contexts awaits future work” (2008, p. 21).

Douglas, Heather (2008). Science, Policy and the Value-Free Ideal. *University of Pittsburgh Press*.

Optimal Steering: Anticipate Performative Effects

“[P]roactively embrace the causal impacts of prediction to find performatively ‘optimal’ decision rules. Importantly, optimality could entail the desire to *forecast* future outcomes accurately, as well as to *steer* data distributions towards socially desirable targets” (Perdomo, 2023).

Perdomo, Juan C. (2023). *Performative Prediction: Theory and Practice*. PhD Thesis. Electrical Engineering and Computer Sciences University of California, Berkeley.

Optimal Steering: Anticipate Performative Effects

"[P]roactively embrace the causal impacts of prediction to find performatively 'optimal' decision rules. Importantly, optimality could entail the desire to *forecast* future outcomes accurately, as well as to *steer* data distributions towards socially desirable targets" (Perdomo, 2023).

Perdomo, Juan C. (2023). *Performative Predict* is. Electrical Engineering and Computer Sciences University

Optimal Steering

But how would this look in prediction-allocation contexts?

- Would we steer the case-worker to make the socially desirable decision?
- How do we decide what the socially desirable decisions are?

Optimal steering flirts with technocratic authoritarianism and objectifies people as bits of “nature” whose behavior can be anticipated (but who are not be reasoned with).

Inductive Risk (in Performative Contexts)

“Certainly we should expect honesty and forthrightness from our scientists. To deliberately deceive decisionmakers or the public in an attempt to steer decisions in a particular direction for self-interested reasons is not morally acceptable. Not only would such a course violate the ideals of honesty central to basic science, but it would violate the basic ideal of democracy, that an elite few should not subvert the will of the many for their own gain” (p. 80).

Douglas, Heather (2008). Science, Policy and the Value-Free Ideal. *University of Pittsburgh Press*.

Optimal Steering

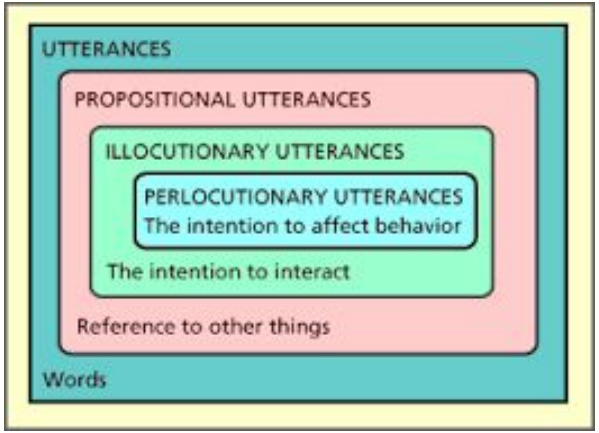
But perhaps the most serious objection is not moral, but methodological.

- What happens when the case-worker realizes she is being steered?
- Do we have to anticipate her reaction to this as well?

It is not plausible that performative effects are sufficiently stable in social contexts to be able to reliably anticipate them.

Going Perlocutionary

“A speaker can pursue perlocutionary aims only when he deceives his partner concerning the fact that he is acting strategically … as soon as there is a danger that these will be attributed to the speaker as intended results the latter finds it necessary to offer explanations and denials, and if need be, apologies, in order to dispel the impression that these side effects are perlocutionary effects. Otherwise … the other participants will feel deceived and **adopt a strategic attitude in turn**, steering away from action oriented to reaching understanding” (295).



Habermas, Jürgen (1981). *Theorie des Kommunikativen Handels, Band I*. Suhrkamp Verlag.

Decision Focused Learning

Prediction-focused learning factors policy applications into

(1) a **prediction step**, in which you try to predict outcomes as “accurately” as possible, and

(2) an **allocation step**, in which predictions from step (1) are fed into an optimization procedure that outputs the most socially desirable allocation of resources.

Decision Focused Learning

Prediction-focused learning factors policy applications into

(1) a **prediction step**, in which you try to predict outcomes as “accurately” as possible, and

(2) an **allocation step**, in which predictions from step (1) are fed into an optimization procedure that outputs the most socially desirable allocation of resources.

While *perfect* predictions would lead to optimal allocations, it is sometimes more practical to estimate allocation policies “directly”.

Decision Focused Learning

Decision-focused learning favors an “end-to-end” system in which steps (1-2) are repeated until an optimal policy is found

(1) **predictions** are used to arrive at optimal allocations and

(2) errors in allocation are back-propagated to update predictions in part (1).

Decision-Focused Learning: Foundations, State of the Art, Benchmark and Future Opportunities

Jayanta Mandi * JAYANTA.MANDI@KULEUVEN.BE
KU Leuven, Belgium

James Kotary † JK4PN@VIRGINIA.EDU
University of Virginia, USA

Senne Berden SENNE.BERDEN@KULEUVEN.BE
KU Leuven, Belgium

Maxime Mulamba MAXIME.MULAMBA@VUB.BE
Vrije Universiteit Brussel, Belgium

Víctor Bucarey VICTOR.BUCAREY@UOH.CL
Universidad de O'Higgins, Chile

Tias Guns TIAS.GUNS@KULEUVEN.BE
KU Leuven, Belgium

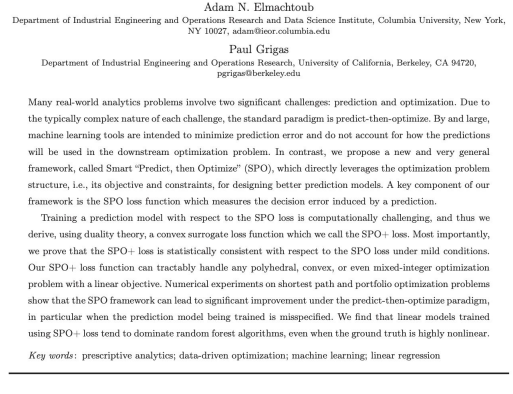
Ferdinando Fioretto FIORETTO@VIRGINIA.EDU
University of Virginia, USA

Abstract

Decision-focused learning (DFL) is an emerging paradigm that integrates machine learning (ML) and constrained optimization to enhance decision quality by training ML models in an end-to-end system. This approach shows significant potential to revolutionize combinatorial decision-making in real-world applications that operate under uncertainty, where estimating unknown parameters within decision models is a major challenge. This paper presents a comprehensive review of DFL, providing an in-depth analysis of both gradient-based and gradient-free techniques used to combine ML and constrained optimization. It evaluates the strengths and limitations of these techniques and includes an extensive empirical evaluation of eleven methods across seven problems. The survey also offers insights into recent advancements and future research directions in DFL.

Code and benchmark: `https://github.com/PredOpt/predopt-benchmarks`

1. Introduction



Smart “Predict, then Optimize”

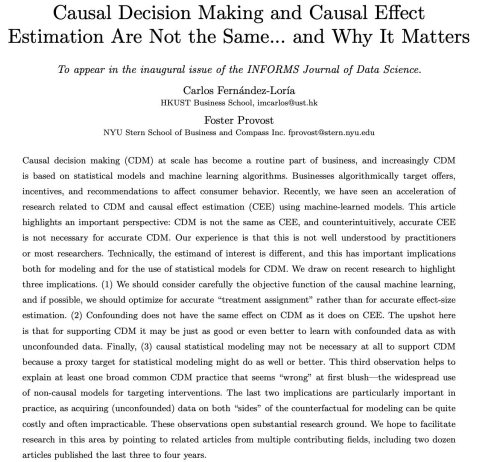
Adam N. Elmachtoub
Department of Industrial Engineering and Operations Research and Data Science Institute, Columbia University, New York, NY 10027, adam@ieor.columbia.edu

Paul Grigas
Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720, pgrigas@berkeley.edu

Many real-world analytics problems involve two significant challenges: prediction and optimization. Due to the typically complex nature of each challenge, the standard paradigm is predict-then-optimize. By and large, machine learning tools are intended to minimize prediction error and do not account for how the predictions will be used in the downstream optimization problem. In contrast, we propose a new and very general framework, called Smart “Predict, then Optimize” (SPO), which directly leverages the optimization problem structure, i.e., its objective and constraints, for designing better prediction models. A key component of our framework is the SPO loss function which measures the decision error induced by a prediction.

Training a prediction model with respect to the SPO loss is computationally challenging, and thus we derive, using duality theory, a convex surrogate loss function which we call the SPO+ loss. Most importantly, we prove that the SPO+ loss is statistically consistent with respect to the SPO loss under mild conditions. Our SPO+ loss function can tractably handle any polyhedral, convex, or even mixed-integer optimization problem with a linear objective. Numerical experiments on shortest path and portfolio optimization problems show that the SPO framework can lead to significant improvement under the predict-then-optimize paradigm, in particular when the prediction model being trained is misspecified. We find that linear models trained using SPO+ loss tend to dominate random forest algorithms, even when the ground truth is highly nonlinear.

Key words: prescriptive analytics; data-driven optimization; machine learning; linear regression



Causal Decision Making and Causal Effect Estimation Are Not the Same... and Why It Matters

To appear in the inaugural issue of the INFORMS Journal of Data Science.

Carlos Fernández-Loría
HKUST Business School, imcarlos@ust.hk

Foster Provost
NYU Stern School of Business and Compass Inc. fprovost@stern.nyu.edu

Causal decision making (CDM) at scale has become a routine part of business, and increasingly CDM is based on statistical models and machine learning algorithms. Businesses algorithmically target offers, incentives, and recommendations to affect consumer behavior. Recently, we have seen an acceleration of research related to CDM and causal effect estimation (CEE) using machine-learned models. This article highlights an important perspective: CDM is not the same as CEE, and counterintuitively, accurate CEE is not necessary for accurate CDM. Our experience is that this is not well understood by practitioners or most researchers. Technically, the estimand of interest is different, and this has important implications both for modeling and for the use of statistical models for CDM. We draw on recent research to highlight three implications. (1) We should consider carefully the objective function of the causal machine learning, and if possible, we should optimize for accurate “treatment assignment” rather than for accurate effect-size estimation. (2) Confounding does not have the same effect on CDM as it does on CEE. The upshot here is that for supporting CDM it may be just as good or even better to learn with confounded data as with unconfounded data. Finally, (3) causal statistical modeling may not be necessary at all to support CDM because a proxy target for statistical modeling might do as well or better. This third observation helps to explain at least one broad common CDM practice that seems “wrong” at first blush—the widespread use of non-causal models for targeting interventions. The last two implications are particularly important in practice, as acquiring (unconfounded) data on both “sides” of the counterfactual for modeling can be quite costly and often impracticable. These observations open substantial research ground. We hope to facilitate research in this area by pointing to related articles from multiple contributing fields, including two dozen articles published the last three to four years.

Decision Focused Learning

Decision-focused learning favors an “end-to-end” system in which steps (1-2) are repeated until an optimal policy is found

(1) **predictions** are used to arrive at optimal allocations and

(2) errors in allocation are back-propagated to update predictions in part (1).

Decision-focused learners argue that (1) there is no independent predictive context in which only the value of predictive accuracy reigns and (2) the only relevant prediction errors are the ones that affect allocation decisions. Optimize for **optimal allocation**, rather than accurate estimation of treatment effects.

Decision Focused Learning

Decision-focused learning favors an "end-to-end" system in which steps (1-2) are repeated until an optimal policy is found

(1) **predictions** are used to arrive at optimal allocations and

(2) errors in allocation are back-propagated to update predictions in part (1).

Decision-focused learners argue that (1) there is no independent predictive context in which only the value of predictive accuracy reigns and (2) the only relevant prediction errors are the ones that affect allocation decisions. Optimize for **optimal allocation**, rather than accurate estimation of treatment effects.

Pragmatic encroachment in ML! Managing inductive risks is more important than predictive "accuracy."

Decision Focused Learning: Difficulties

DFL faces many difficulties:

(1) Is it even possible to back-propagate errors in policy settings? Where do you get reliable signals of policy error?

(2) Predictions are finely tuned to a policy goal and cannot be re-used in other policy contexts.

(3) ML engineers are empowered to make normative decisions about appropriate (surrogate) local justice principles. Backprop favors *differentiable* loss functions.

(4) Cannot be used to adjudicate between different policies!

Decision Focused Learning: Difficulties

DFL faces many difficulties:

(1) Is it even possible to back-propagate errors in policy settings? Where do you get reliable signals of policy error?

(2) Predictions are finely tuned to a policy goal and cannot be re-used in other policy contexts.

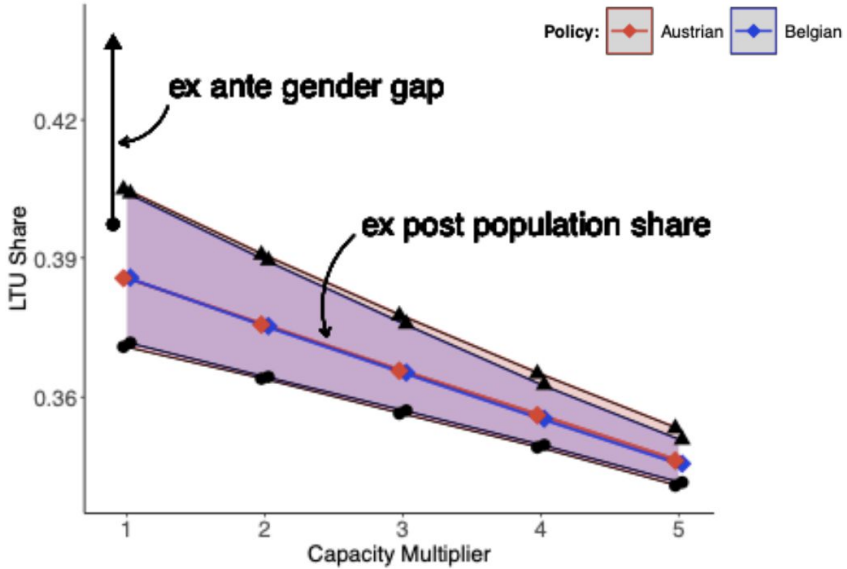
(3) ML engineers are empowered to make normative decisions about appropriate (surrogate) local justice principles. Backprop favors *differentiable* loss functions.

(4) Cannot be used to adjudicate between different policies!

Decision Focused Learning: Difficulties

If predictions of long-term unemployment are permeated by the values of hawks, or of doves, then these predictions cannot be used to forecast the consequences of hawkish/dovish policies and adjudicate disputes between them.

And what if **both** could have been satisfied?



Practical Rationality

“Weber differentiates the concept of practical rationality from the three perspectives of *employing means*, *setting ends*, and *being oriented to values*.

- The instrumental rationality of an action is measured by effective planning of the application of means for given ends;
- the rationality of choice of an action is measured by the correctness of the calculation of ends in light of precisely conceived values, available means and boundary conditions;
- and the normative rationality of an action is measured by the unifying, systematizing power and penetration of the value standards and the principles that underlie action preferences.” (p. 172)

Habermas, Jürgen (1981). *Theorie des Kommunikativen Handels, Band I*. Suhrkamp Verlag.

Communicative Rationality

“In the context of communicative action, only those persons count as responsible who, as members of a communication-community, can orient their actions to **intersubjective recognized validity claims**. …

A greater degree of cognitive-instrumental rationality produces a greater independence from limitations imposed by the environment on the self-assertion of subjects acting in a goal-directed manner.

A greater degree of communicative rationality expands—within a communication-community—the scope for unconstrained coordination of actions and consensual resolution of conflicts.”

Habermas, Jürgen (1981). *Theorie des Kommunikativen Handels, Band I*. Suhrkamp Verlag.

Du Bois' Democratic Defence of the Value-Free Ideal

"... too closely pairing scientific argument with reform efforts in the past has resulted in 'closely connecting social investigation with a good deal of groundless assumption and humbug in the popular mind.' This can be seen as raising the worry that 'philanthropists and statesman' must be willing and able, in light of public opinion, to make use of scientifically acquired knowledge. They cannot do this in a democracy if the public think their results untrustworthy or are for whatever reason unwilling to act (or vote in those who would have them act) upon it …. " (p. 2233)

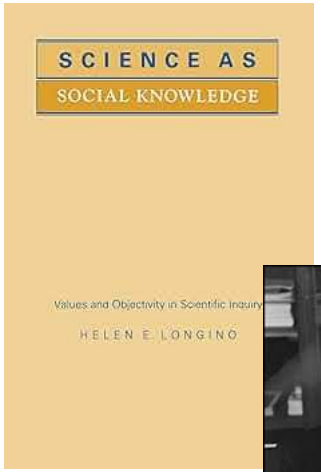
Bright, Liam Kofi (2018). "Du Bois' democratic defence of the value free ideal." *Synthese*, 195(5), 2227-2245.

Du Bois' Democratic Defence of the Value-Free Ideal

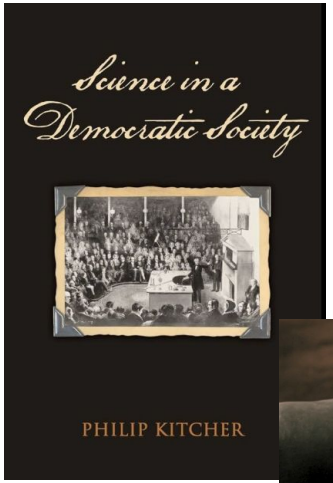
"... [I]f one's mediate aim for science is to guide democratic policy making, one will have reason to insist that scientists' immediate aim be **pure truth seeking.** " (p. 2233)

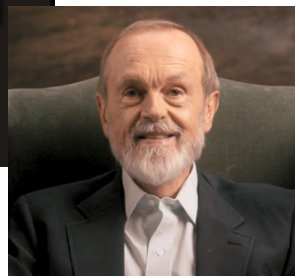
Bright, Liam Kofi (2018). "Du Bois' democratic defence of the value free ideal." *Synthese*, 195(5), 2227-2245.

Discourse-Ethical Philosophy of Science



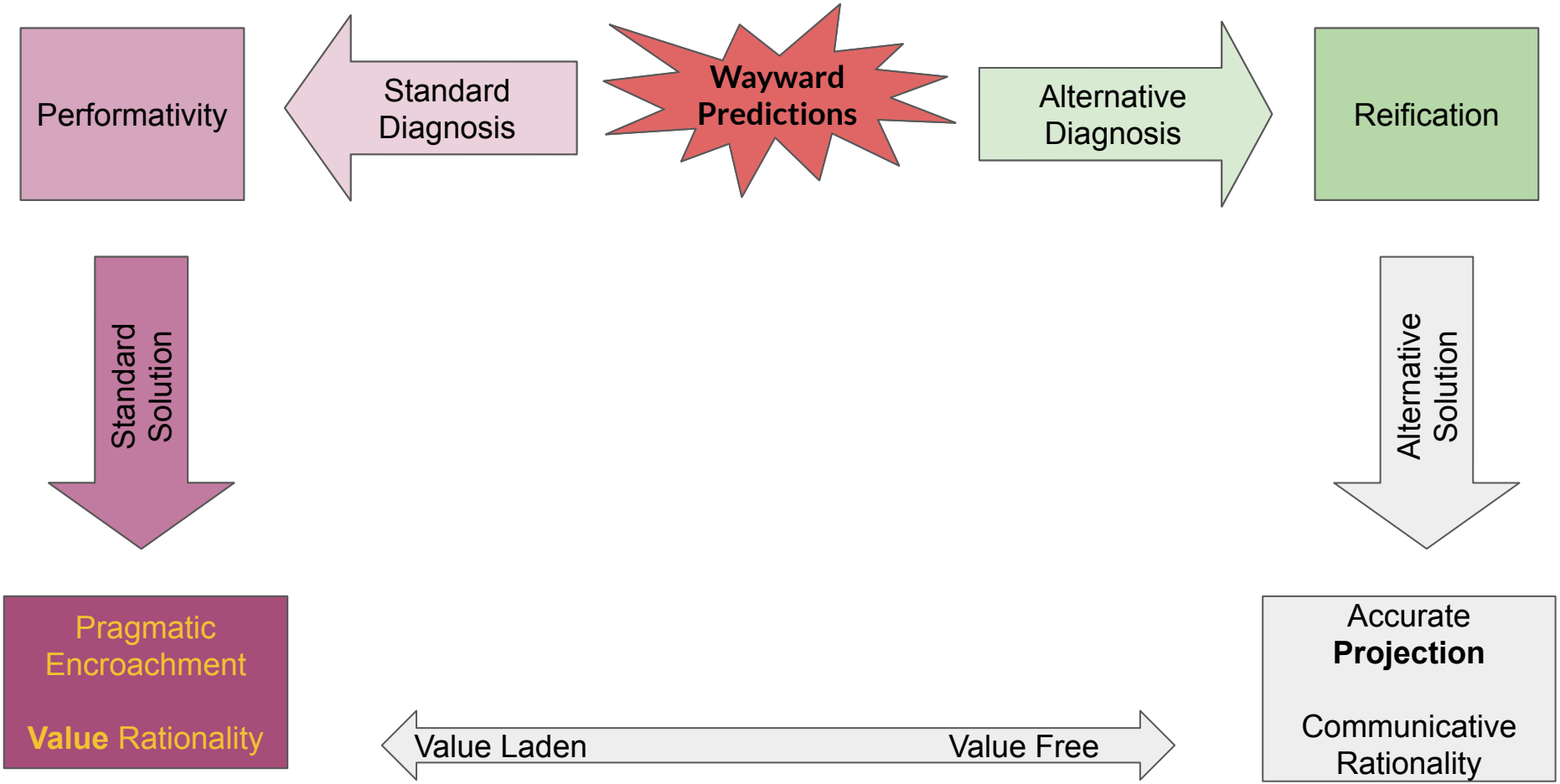
1990



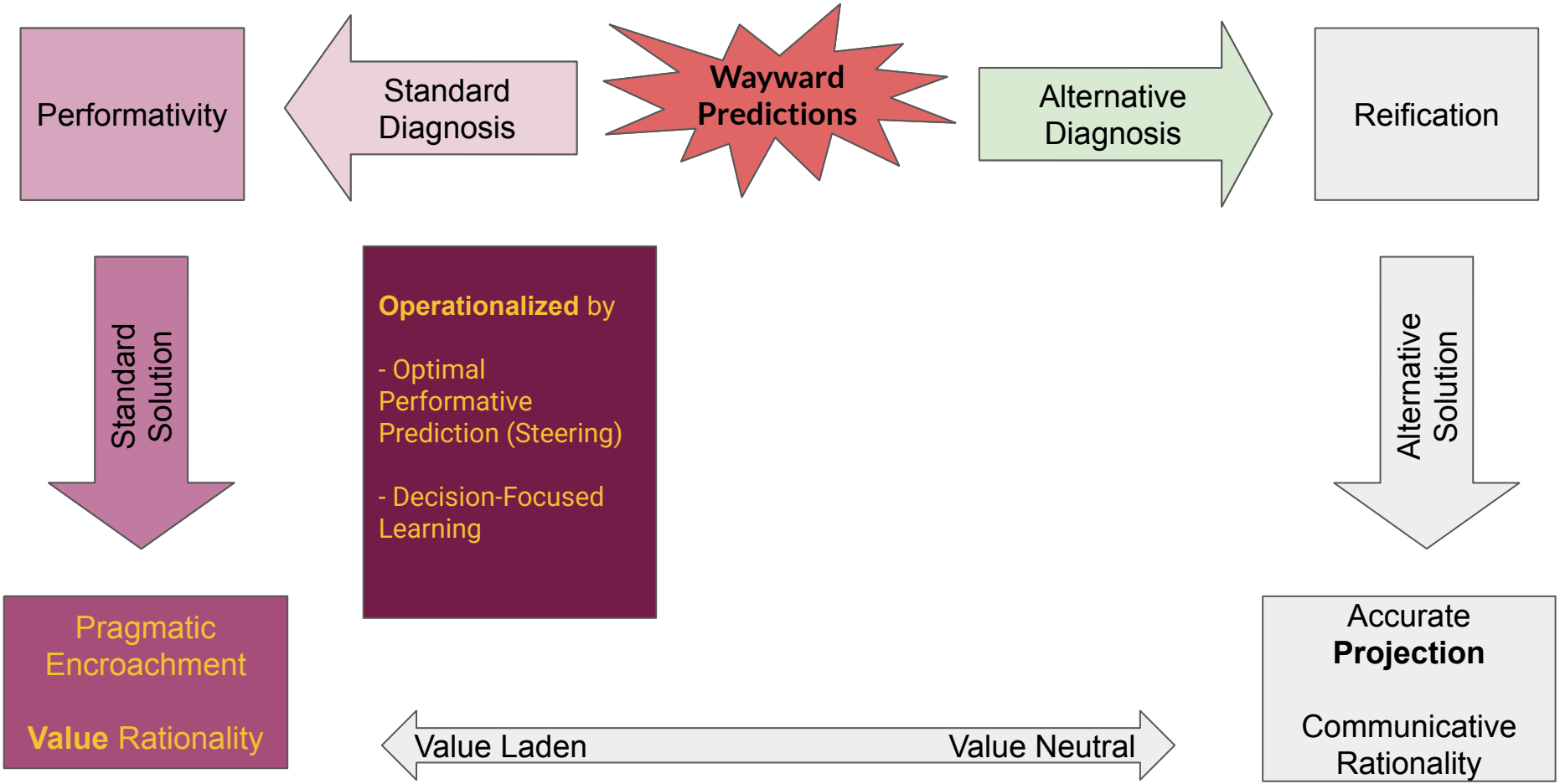


2011

Wayward Predictions



Wayward Predictions



Performativity: Micro

Recall: We are interested in **micro**-performativity, in which a local prediction about an individual causally influences that individuals outcomes. These effects are typically **intended** and (somewhat) easier to anticipate (Zezulka and Genin, 2023; Kim and Perdomo, 2022).

Kim, Michael P., and Juan C. Perdomo. (2022) "Making decisions under outcome performativity."

Zezulka, S. and Genin, K. (2023). “Performativity and Prospective Fairness” *Fairness Through the Lens of Time*, NeurIPS 2023.

Performativity: Micro

Recall: We are interested in **micro**-performativity, in which a local prediction about an individual causally influences that individuals outcomes. These effects are typically **intended** and (somewhat) easier to anticipate (Zezulka and Genin, 2023; Kim and Perdomo, 2022).

Is this intended, foreseeable performativity really a problem?

Kim, Michael P., and Juan C. Perdomo. (2022) "Making decisions under outcome performativity."

Zezulka, S. and Genin, K. (2023). “Performativity and Prospective Fairness” *Fairness Through the Lens of Time*, NeurIPS 2023.

Performativity: Micro

Recall: We are interested in **micro**-performativity, in which a local prediction about an individual causally influences that individuals outcomes. These effects are typically **intended** and (somewhat) easier to anticipate (Zezulka and Genin, 2023; Kim and Perdomo, 2022).

Is this intended, foreseeable performativity really a problem? No!

Kim, Michael P., and Juan C. Perdomo. (2022) "Making decisions under outcome performativity."

Zezulka, S. and Genin, K. (2023). “Performativity and Prospective Fairness” *Fairness Through the Lens of Time*, NeurIPS 2023.

Performativity: Micro

Recall: We are interested in **micro**-performativity, in which a local prediction about an individual causally influences that individuals outcomes. These effects are typically **intended** and (somewhat) easier to anticipate (Zezulka and Genin, 2023; Kim and Perdomo, 2022).

Is this intended, foreseeable performativity really a problem? No! The problem is that *predictions* are ambiguous in social contexts. (And often delivered with the wrong illocutionary force.)

Kim, Michael P., and Juan C. Perdomo. (2022) "Making decisions under outcome performativity."

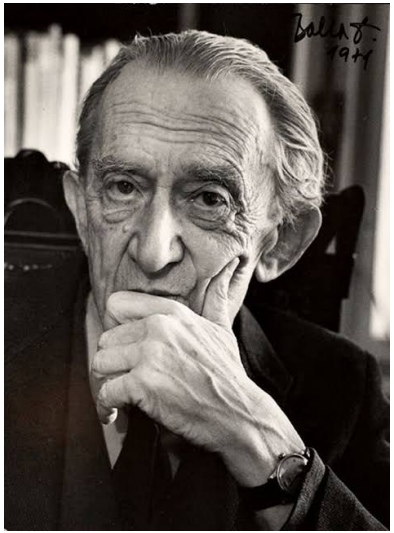
Zezulka, S. and Genin, K. (2023). “Performativity and Prospective Fairness” *Fairness Through the Lens of Time*, NeurIPS 2023.

Reification

Wayward prediction is caused by *reification*: social-statistical regularities are presented as ‘natural laws’ to which we can *accommodate* ourselves, but which we cannot alter. They obscure the fact that these outcomes are, to a significant degree, *under our control.*

Reification

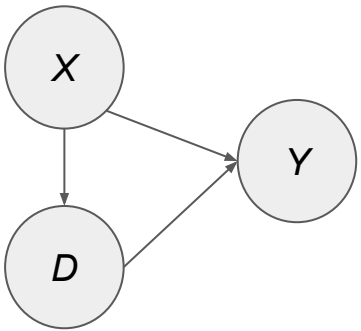
“[M]an’s own activity … becomes something objective and independent of him, something that controls him by virtue of of an autonomy alien to man. … The laws governing these objects are indeed gradually discovered by man, but even so they confront him as invisible forces that generate their own power. The individual can **use** his knowledge of these laws to his own advantage, but he is not able to **modify** the process by his own activity.”



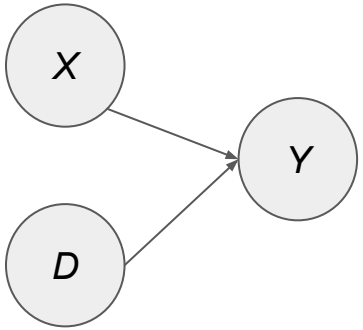
Lukács, György (1923). “Reification and the Consciousness of the Proletariat.” in *History and Class Consciousness*, MIT Press.

Target Specification Bias

Risk scores are estimated by data generated by the actual scenario (left), but decision makers need predictions under the counterfactual scenario (right).



Actual scenario

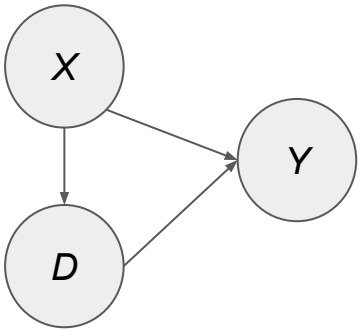


Counterfactual scenario

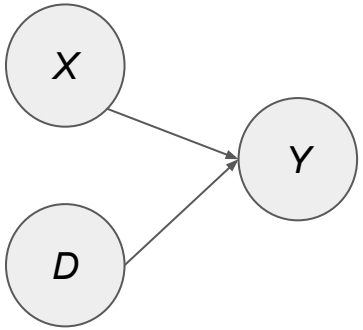
Tal, Eran. (2023) Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare.

Target Specification Bias

Risk scores predict outcomes in part by *guessing* what treatment someone will receive.



Actual scenario

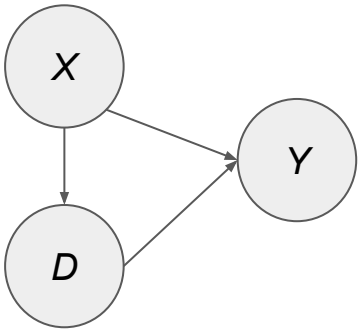


Counterfactual scenario

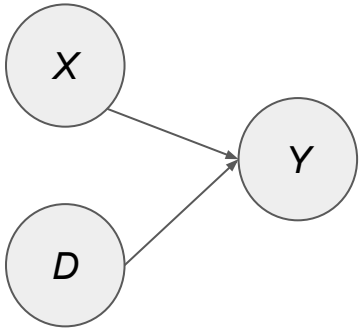
Tal, Eran. (2023) Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare.

Target Specification Bias

Risk scores, as normally understood, only confound decision making — they should play no role when target specification bias is significant.



Actual scenario



Counterfactual scenario

Tal, Eran. (2023) Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare.

Projections (rather than Forecasts)

“[Y]ou can think of **explanations** as providing an account of how something happened, **projections** as providing predictions about what would happen under certain hypothetical conditions, and **forecasts** as indicating what can be expected to happen.”

Lukács, György (1923). “Reification and the Consciousness of the Proletariat.” in *History and Class Consciousness*, MIT Press.

Projections (rather than Forecasts)

“[Y]ou can think of **explanations** as providing an account of how something happened, **projections** as providing predictions about what would happen under certain hypothetical conditions, and **forecasts** as indicating what can be expected to happen.”

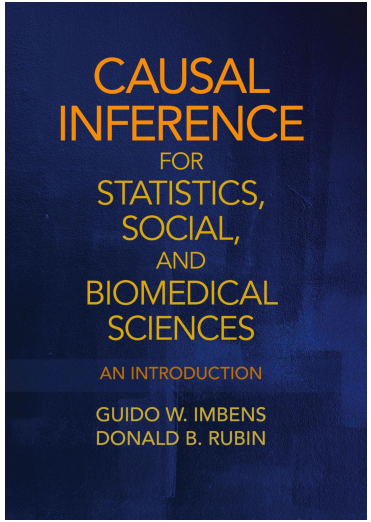
Lukács, György (1923). “Reification and the Consciousness of the Proletariat.” in *History and Class Consciousness*, MIT Press.

Potential Outcomes

In the potential outcome framework, $Y_i(t)$ represents the outcome for individual i if she ***were to receive*** treatment t.

The fundamental projection problem is to infer $\hat{Y}_i(t)$ for every individual and treatment (e.g., via double-robust machine learning).

Potential Outcomes are **less reified** than forecasts because they 'wear on their face' the fact that outcomes are, at least in part, an expression of our decisions.



Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal.

Michael C. Knaus. 2022. Double machine learning-based programme evaluation under unconfoundedness

Potential Outcomes

We claim: **accurately** estimating potential outcomes solves the problem of wayward prediction.

Rather than forecasting mortality, present physicians with projections of mortality under inpatient/outpatient treatment.

Rather than forecasting unemployment, present case-workers with projections of unemployment under the different program options.

Potential Outcomes

We claim: **accurately** estimating potential outcomes solves the problem of wayward prediction.

Rather than forecasting mortality, present physicians with projections of mortality under inpatient/outpatient treatment.

Rather than forecasting unemployment, present case-workers with projections of unemployment under the different program options.

But why shouldn't pragmatic encroachment arguments now be directed at potential outcomes?

Decision-Supportive Projections

Say that a projection is **decision-supportive** for A if it allows agent A to accurately evaluate the policy options in light of her goals and values.

Techniques from the pragmatic encroachment family are decision-supportive (value-rational).

Discourse-Supportive Projections

In discourse, several agents discuss collectively which of a number of policy options is best all-things-considered. In the course of discussion, agents must be able to sympathetically identify with the goals and values of others, even if they ultimately reject them in favor of their own. Ideally, one policy option is best from all perspectives, but if this is not the case, it is also important to see whether the best policy from the perspective of A is not also acceptable from the perspective of B.

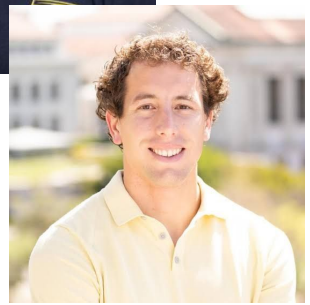
Discourse-Supportive Projections

Say that a projection is **discourse-supportive** for $A_1, A_2 \ldots, A_n$ if it allows each agent to accurately evaluate the policy options in light of her own goals and values as well as the goals and values **of the other participants in the discourse**.

Accurate estimates of potential outcomes are discourse-supportive (communicatively rational).

Performative Omniprediction

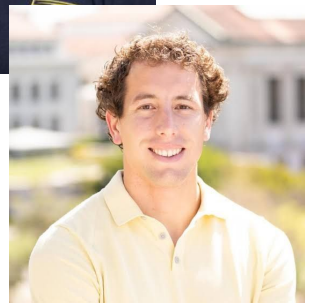
“we have assumed that the system designer is able to adequately encode the overarching goals of prediction into a single, fixed loss function ℓ that we then optimize via the performative risk. That is, we assume that this normative task of translating the subjective goals of prediction into a concrete, objective mathematical object has been resolved a priori. … ”



Kim, Michael and Perdomo, Juan C. (2022). “Making Predictions under Outcome Performativity.”

Performative Omniprediction

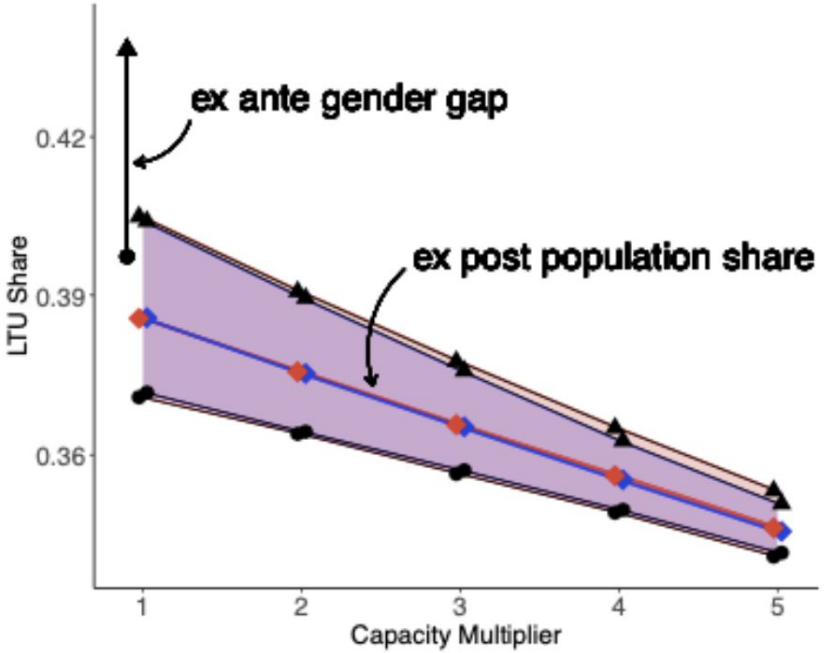
“In this chapter, we attempt to find technical solutions that directly embrace the multiplicity of objectives in performative prediction. … [W]e introduce the concept of a performative omnipredictor and illustrate how these solutions can be learned efficiently. Intuitively, a performative omnipredictor is a single predictive model that is simultaneously performatively optimal for many, diverse objectives. By diverse, we mean that these **omnipredictors can be used to generate optimal predictions for qualitatively different, and possibly contradictory goals.** ”



Kim, Michael and Perdomo, Juan C. (2022). “Making Predictions under Outcome Performativity.”

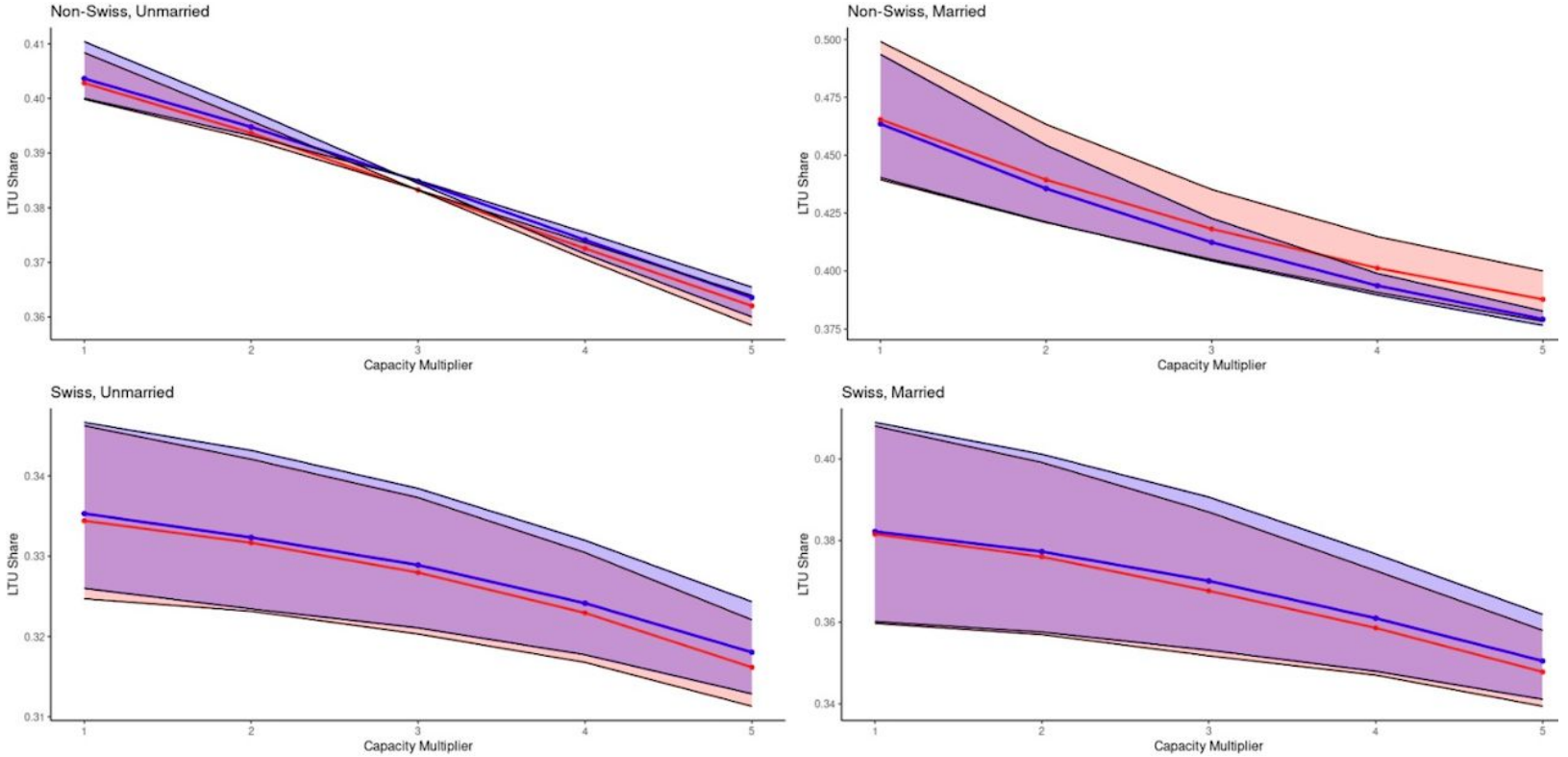
Hawks and Doves

We find no efficiency gains in neglecting those at highest risk: Austrian prioritization is slightly worse overall and creates larger gender gaps.

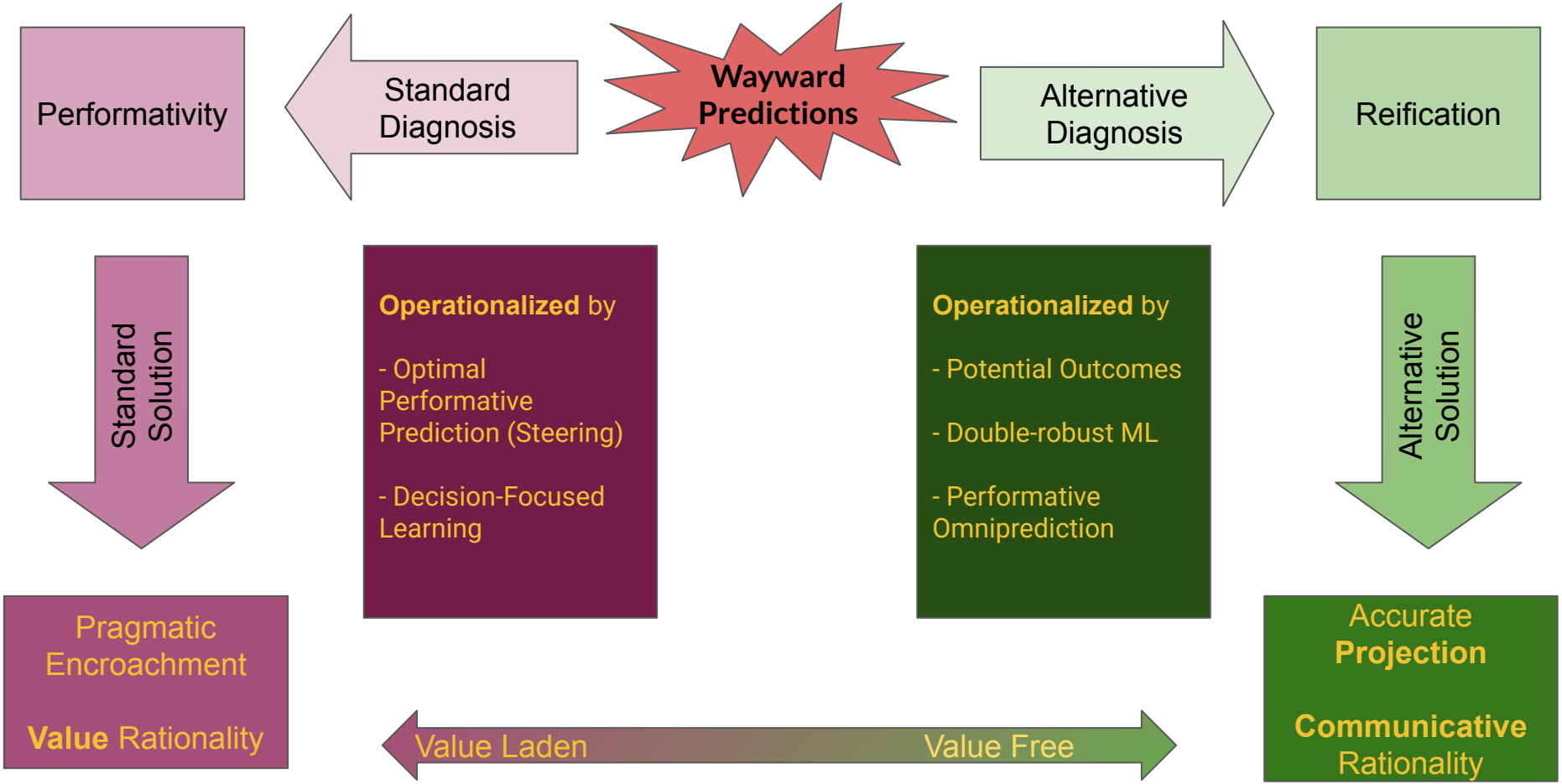


Hawks and Doves

Belgian policy narrows gender gaps among the least advantaged (married, non-citizens).



Wayward Predictions



Thank You!

From the fair distribution of predictions ...

Algorithmic fairness focuses on the distribution of ***predictions*** at the time of ***training***, rather than the distribution of ***social goods*** induced by ***deploying*** an algorithm in a concrete policy context.

… to the fair distribution of social goods.

Algorithmic fairness focuses on the distribution of ***predictions*** at the time of ***training***, rather than the distribution of ***social goods*** induced by ***deploying*** an algorithm in a concrete policy context.

The distribution of	what?	when?	by whom?
Received View	Predictions.	Right after training, before deployment.	Data scientists.
Proposed View	Social goods e.g., jobs, freedoms, spots in schools and universities, social esteem.	*After* deployment.	Data scientists, firms, public agencies, universities, etc.

… to the fair distribution of social goods.

This change in perspective makes algorithmic fairness **continuous** with the egalitarian tradition in distributive justice.

The distribution of	what?	when?	by whom?
Received View	Predictions.	Right after training, before deployment.	Data scientists.
Proposed View	Social goods e.g., jobs, freedoms, spots in schools and universities, social esteem.	*After* deployment.	Data scientists, firms, public agencies, universities, etc.

Fair ML: The Fundamental Question

Will deploying an algorithm in some concrete social context **reproduce** or **exacerbate** the inequalities in the distribution of social goods reflected in their training data?

Widely cited as the ***motivation*** for AI fairness. However, the methodological solutions developed by researchers in algorithmic fairness are, surprisingly, **ill-suited** for answering this fundamental question.

Risk Assessment and Public Employment Services

The algorithm takes as **input**

- the education and employment history (*X*),
- and gender (*A*)

of a recently unemployed person, and **outputs** a risk score (*R*) of long-term unemployment (*Y*).

On the basis of the risk score (*R*), a case-worker assigns the person to some labor-market program (*D)* that is causally relevant for their employment prospects *(Y*) .

From the Fair Distribution of Predictions to the Fair Distribution of Social Goods: Evaluating the Impact of Fair Machine Learning on Long-Term Unemployment*

Sebastian Zezulka
University of Tübingen
sebastian.zezulka@uni-tuebingen.de

Konstantin Genin
University of Tübingen
konstantin.genin@uni-tuebingen.de

June 2024

Abstract

Deploying an algorithmically informed policy is a significant intervention in society. Prominent methods for algorithmic fairness focus on the distribution of *predictions* at the time of *training*, rather than the distribution of *social goods* that arises *after* deploying the algorithm in a specific social context. However, requiring a 'fair' distribution of predictions may undermine efforts at establishing a fair distribution of social goods. First, we argue that addressing this problem requires a notion of *prospective fairness* that anticipates the change in the distribution of social goods *after* deployment. Second, we provide formal conditions under which this change is identified from pre-deployment data. That requires accounting for different kinds of performative effects. Here, we focus on the way predictions change policy decisions and, consequently, the causally downstream distribution of social goods. Throughout, we are guided by an application from public administration: the use of algorithms to predict who among the recently unemployed will remain unemployed in the long term and to target them with labor market programs. Third, using administrative data from the Swiss public employment service, we simulate how such algorithmically informed policies would affect gender inequalities in long-term unemployment. When risk predictions are required to be 'fair' according to statistical parity and equality of opportunity, targeting decisions are less effective, undermining efforts to both lower overall levels of long-term unemployment and to close the gender gap in long-term unemployment.

*Sebastian Zezulka and Konstantin Genin. 2024. From the Fair Distribution of Predictions to the Fair Distribution

Risk Assessment and Public Employment Services

The risk score may support a number of different policies.

- In Belgium: individuals at high risk of long-term unemployment are **prioritized** (Desiere and Struyven, 2020).

- In Austria: risk scores classify the recent unemployed into those with (i) good prospects in the next six months; (ii) bad prospects in the next two years; and (iii) everyone else. Support measures **target** the third group, while offering **only limited support** to the first and second group (Allhutter et al., 2020).

Allhutter, D., Cech, F., Fischer, F., Grill G, and Mager, A. (2020) Algorithmic profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective.

Sam Desiere and Ludo Struyven. Using Artificial Intelligence to classify Jobseekers: The Accuracy-Equity Trade-off. Journal of Social Policy, 50(2):367–385, 2020.

Risk Assessment and Public Employment Services

Advocates of the Austrian policy argue in terms of *efficiency.*

Risk Assessment and Public Employment Services

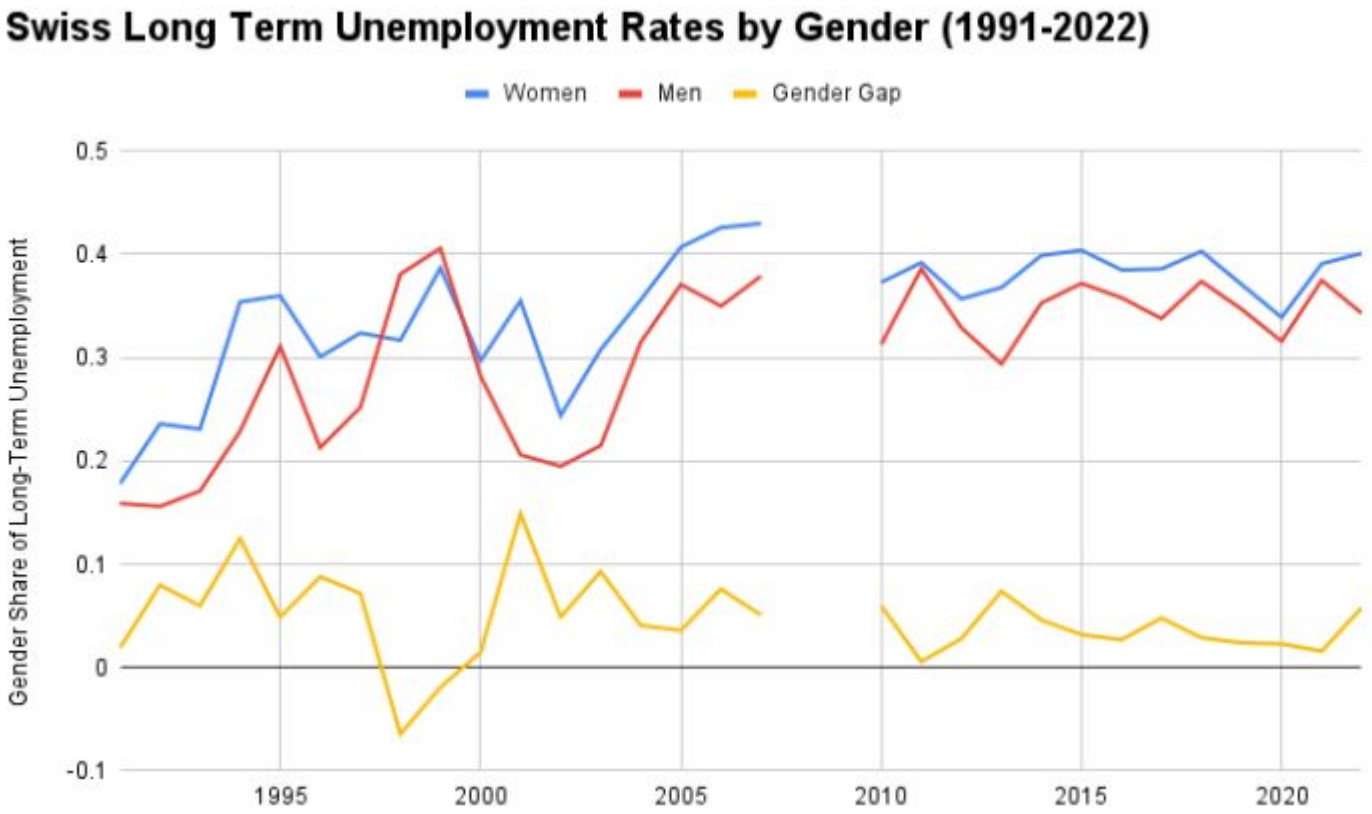
Advocates of the Austrian policy argue in terms of *efficiency.*

But what if the highest risk are simply more likely to get the least effective programs?

Risk Assessment and Public Employment Services

Critics of the Austrian plan worry about exacerbating long-standing structural inequalities in the labor market.

The Gender Reemployment Gap: Switzerland



Algorithmic Fairness to the Rescue?

At first approximation: fairness notions are what you can express with the sensitive attribute (A), the risk score (R), the outcome (Y) and conditional probability.

- **Demographic Parity:** $A \perp R$

  In expectation, men and women should get the same risk scores.

- **Sufficiency:** $A \perp Y \mid R$

  In expectation, people with the same score should have the same outcome, regardless of gender.

- **Separation:** $A \perp R \mid Y$

  In expectation, people with the same outcomes should have the same score, regardless of gender.

Algorithmic Fairness to the Rescue?

At first approximation: fairness notions are what you can express with the sensitive attribute (A), the risk score (R), the outcome (Y) and conditional probability.

- **Demographic Parity:** $A \perp R$

  In expectation, men and women should get the same risk scores.

- **Sufficiency:** $A \perp Y \mid R$

  In expectation, people with the same score should have the same outcome, regardless of gender.

- **Separation:** $A \perp R \mid Y$

  In expectation, people with the same outcomes should have the same score, regardless of gender.

Algorithmic Fairness: Separation

When outcomes are binary, we can factor **Separation** ($A \perp R \mid Y$) into:

Equal False Negatives: $A \perp R \mid Y=1$

Different genders should have equal rates of false negatives.

Equal False Positives: $A \perp R \mid Y=0$

Different genders should have equal rates of false positives.

Algorithmic Fairness: Separation

When outcomes are binary, we can factor **Separation** ($A \perp R \mid Y$) into:

Equal Opportunity: $A \perp R \mid Y=1$

Different genders should have equal rates of false negatives.

Equal False Positives: $A \perp R \mid Y=0$

Different genders should have equal rates of false positives.

Algorithmic Fairness: Separation

So why not Separation?

The distribution of	what?	**when?**	by whom?
Received View	Predictions.	**Right after training, before deployment.**	Data scientists.
Proposed View	Social goods e.g., jobs, freedoms, spots in schools and universities, social esteem.	***After* deployment.**	Data scientists, firms, public agencies, universities, etc.

Separation is Self-Defeating

Mishler and Dalmasso (2020) show that satisfying group-based fairness notions at the time of training **virtually ensures** that they will be violated after deployment:

$$A \perp_{\text{pre}} R \mid Y \text{ entails } A \not\perp_{\text{post}} R \mid Y.$$

Separation is Self-Defeating

In the long-run, Separation can **entrench** systemic inequality (D'Amour et al., 2020).

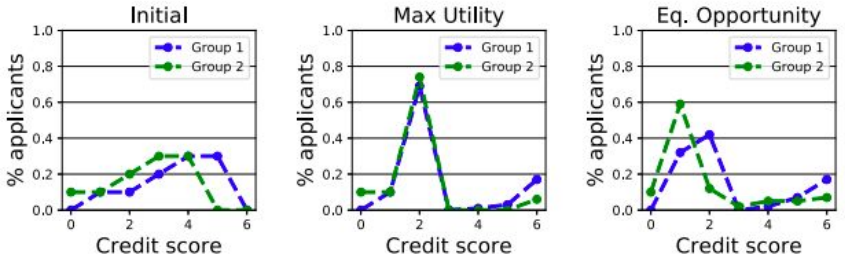


Figure 2: Initial credit score distributions of the two groups (far left) and final states after 20K steps of the environment using a max-util agent (center) and EO agent (right). The credit distributions start with group 2 slightly disadvantaged, but the groups converge to the similar distributions under the max-util agent, while the EO agent maintain unequal credit distributions between groups.

Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. *Fairness Is Not Static: Deeper Understanding of Long Term Fairness via Simulation Studies.* In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, page 525–534

Case Study: Data

We observe ~100k recently unemployed, ages 24-55, registered with the Swiss unemployment service in 2003.

Most are referred either to (1) no program or (2) job search training. Some are referred to (3) computer training; (4) language training; (5) employment programs, (6) personality training or (7) vocational training.

The gender gap in LTU is at 3.9%: 39.7% (men) vs. 43.6% (women).

The citizenship gap in LTU 15.9%: 34.7% (Swiss citizens) vs. 51.5% (non-citizens).

Case Study: Data

	#Obs	LTU	Female (binary)	Age in years	Foreigner (binary)	Employability	Past Income in CHF
Simulation Data	32,148	0.41	0.44	36.8	0.36	1.93	43,461
No program	23,785	0.41	0.43	36.6	0.37	1.92	42,557
Vocational	423	0.28	0.32	37.5	0.32	1.91	49,349
Computer	446	0.24	0.61	38.9	0.20	1.98	43,251
Language	723	0.48	0.54	35.3	0.68	1.83	37,779
Job Search	5,868	0.43	0.44	37.4	0.33	1.98	46,815
Employment	321	0.46	0.43	35.3	0.39	1.84	36,902
Personality	582	0.37	0.35	39.4	0.25	1.93	53,136

Table 1. Descriptive statistics for demographic variables in the simulation data and by observed treatment groups. Long-term unemployment (LTU), Female, and Foreigner are given as shares. Age, Employability, and Past Income are averages. Employability is an ordered variable from low (1) to high (3), assigned by the caseworker. Knaus [46] reports an exchange rate USD/CHF of about 1.3 for 2003.

Case Study: Analysis

1. **Counterfactual Prediction**
   For each individual, estimate the effectiveness of the various programs.

2. **Risk Scores**
   Learn fairness constrained (demographic parity, equal opportunity) and fairness **un**constrained predictors of long-term unemployment.

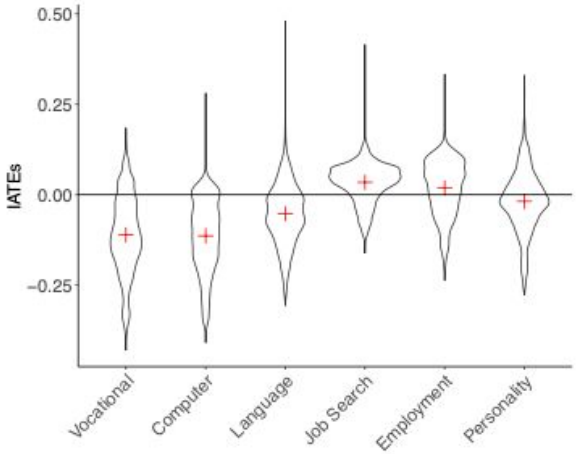
3. **Prioritization**
   For each risk score from part (2), prioritize individuals according to either a Belgian or an Austrian-style scheme.

4. **Allocation**
   For every priority order from step (3), assign unemployed to job programs until capacity constraints are reached.
   - Optimal allocation: assign to the program with the maximum estimated effectiveness;
   - Random allocation: assign uniformly among available programs.

5. **Analysis**
   Forecast the distributional effects of different combinations of choices at steps (1), (3) and (4).

(1) Counterfactual Prediction

For each individual we predict their (counterfactual!) outcomes under each of the seven treatments.*



(a) Individualized Average Treatment Effects.

	ATE	SE	95%-CI
Vocational	-11.12	0.06	[-11.12, -11.12]
Computer	-11.37	0.05	[-11.37, -11.37]
Language	-5.25	0.04	[-5.26, -5.25]
Job Search	3.43	0.03	[3.43, 3.43]
Employment	1.83	0.04	[1.83, 1.83]
Personality	-1.84	0.04	[-1.84, -1.84]

(b) Average Treatment Effects in percentage points, standard errors, and 95% confidence intervals. Negative treatment effects imply a lower risk of becoming long-term unemployed.

Fig. 3. (Individualized) average treatment effects for the six labor market programs. No program serves as baseline.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal.

Michael C. Knaus. 2022. Double machine learning-based programme evaluation under unconfoundedness

John Körtner and Ruben Bach. 2023. Inequality-Averse Outcome-Based Matching.

(2) Risk Scores

Then, we create three risk scores (predictions of *actual* outcomes) for each individual:

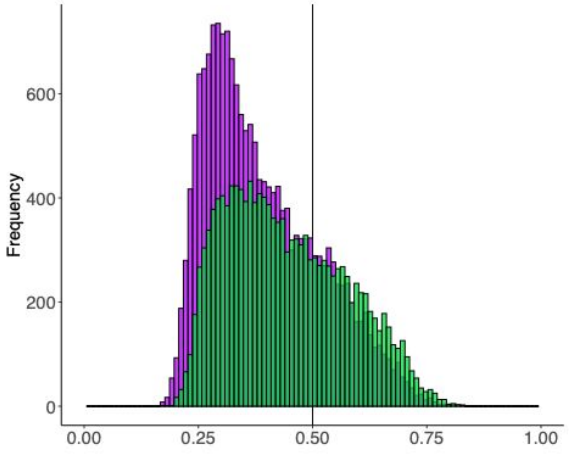
Fairness Unconstrained: We predict LTU using (complexity-penalized) logistic regression.

Demographic Parity: We predict LTU, but encourage risk scores to be independent of gender.

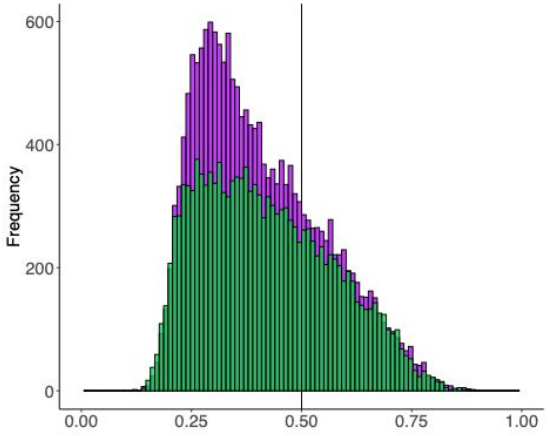
Equal Opportunity: We predict LTU, but encourage risk scores to be independent among the positive class (encourages similar rates of true positives).

(2) Risk Scores

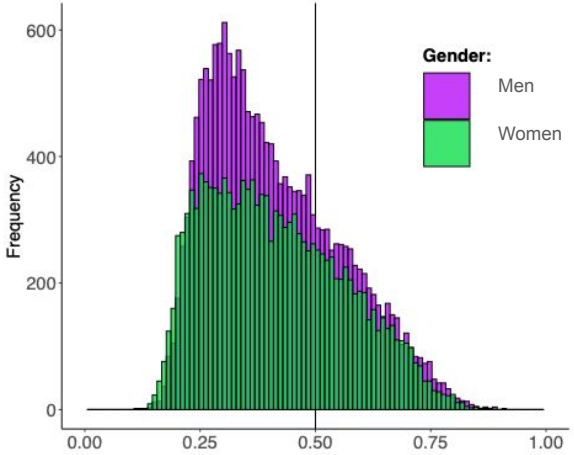
Fairness constraints encourage the distribution of risk to look similar for men and women.



(a) No fairness constraint.



(b) Statistical parity constraint.



(c) Equal opportunity constraint.

(3) Prioritization

For each of the three risk scores from the previous stage, we compile two priority lists modeling the Belgian and Austrian proposals.

Belgian Prioritization: List goes in order of decreasing risk.

Austrian Prioritization: Same, but only for those in the 30 − 70th risk percentiles. The rest are put at the end of the list in random order.

This yields six priority lists, one for each combination of risk score and prioritization scheme.

(4) Allocation

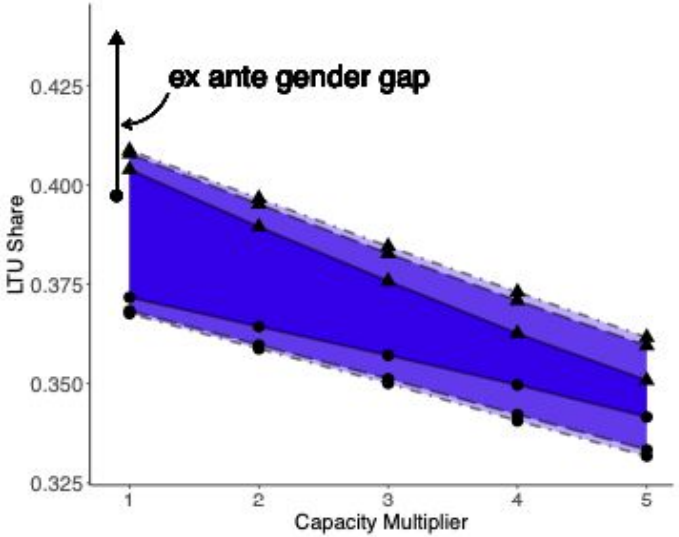
For each of the six lists from the previous stage, we assign individuals to programs in order of priority. Individuals are assigned according to two schemes.

Optimal Assignment: Each person is assigned to the program which is most effective for them and not yet at capacity. This models the best-case scenario in which caseworkers are very good at discerning which program is best for each person.

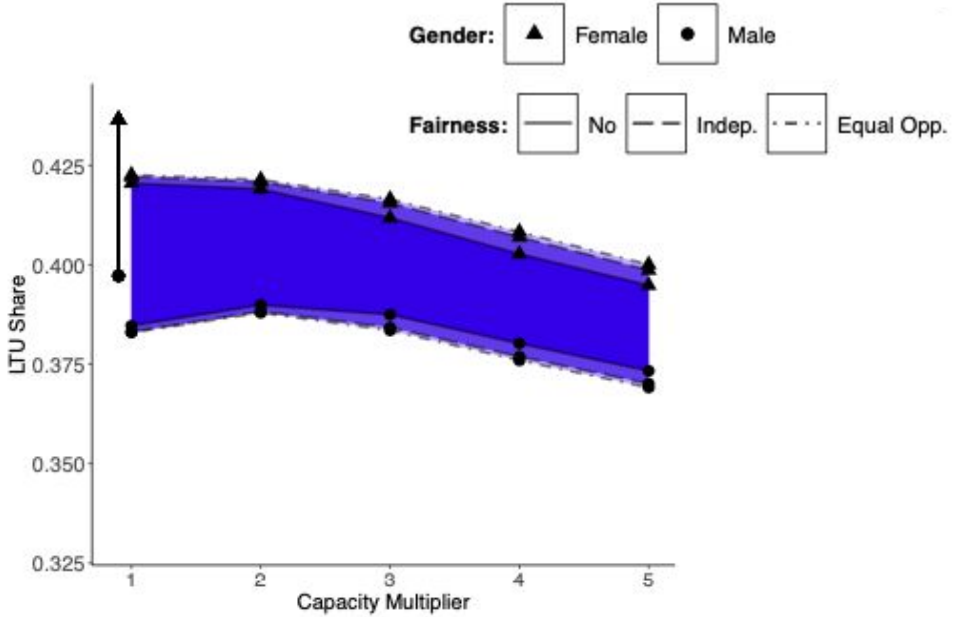
Random Assignment: Each person is assigned by a uniform draw from the available programs. This models the pessimistic scenario in which caseworkers are no better than chance at discerning which program is best for each person.

Fair Predictions vs. Fair Outcomes

No matter what choices are made at other stages, fairness constraints result in larger gender reemployment gaps.



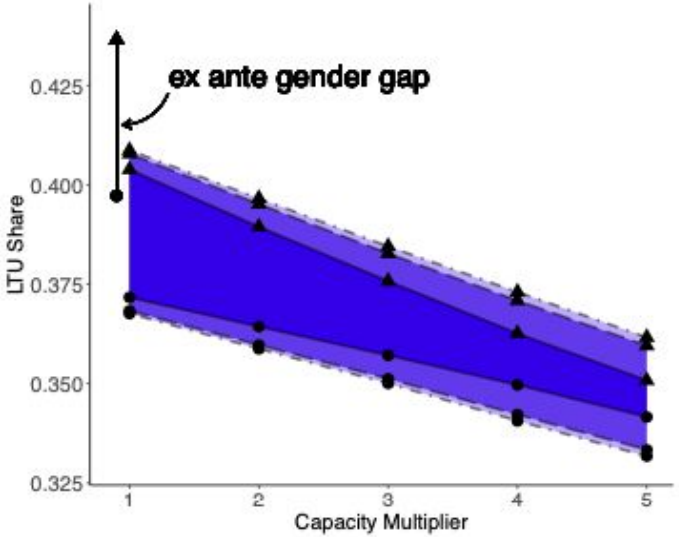
(a) Belgian Prioritization and Optimal Program



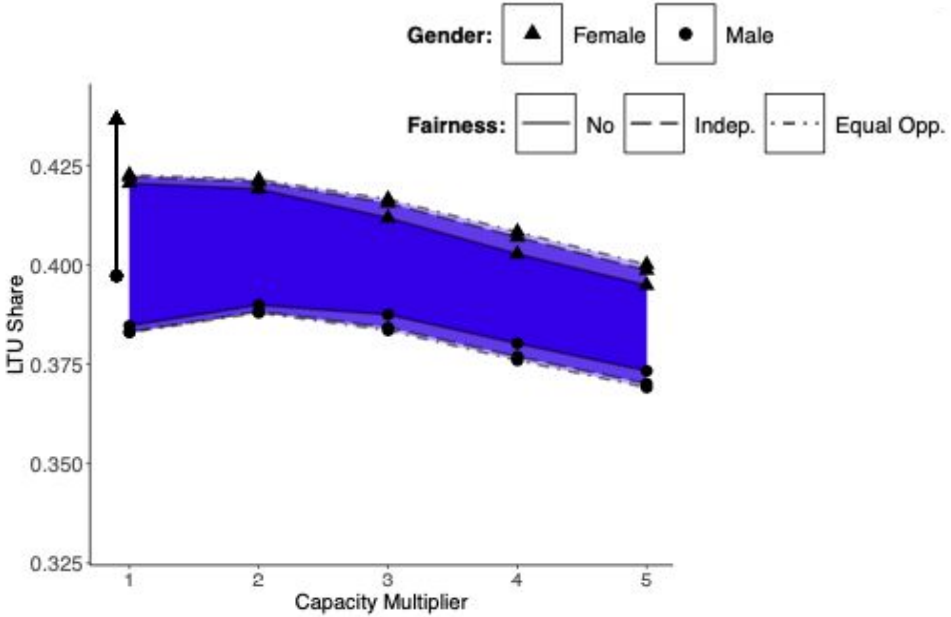
(b) Belgian Prioritization and Random Program

Fairness and Accurate (Counterfactual) Prediction

No matter what choices are made at other stages, individualized assignment results in uniformly better overall outcomes and smaller gender reemployment gaps.



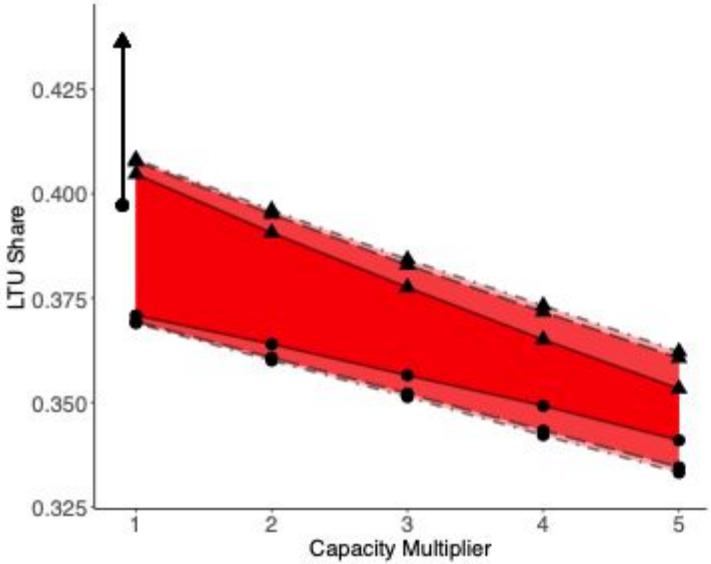
(a) Belgian Prioritization and Optimal Program



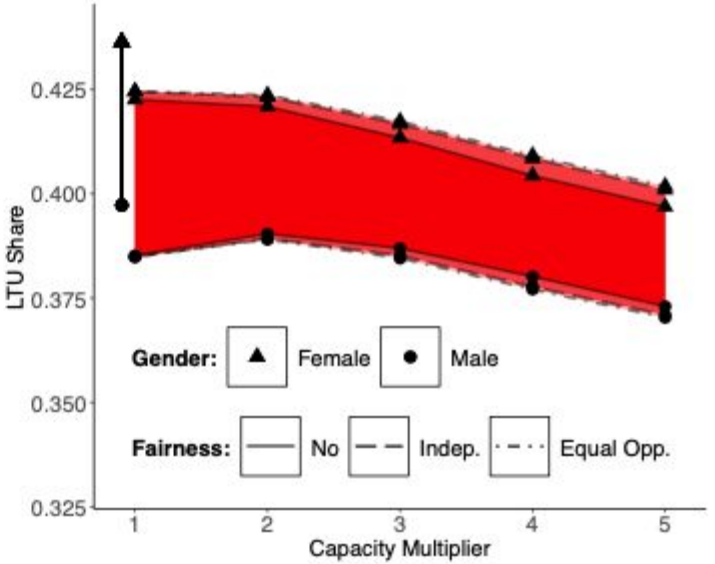
(b) Belgian Prioritization and Random Program

Fair Predictions vs. Fair Outcomes

No matter what choices are made at other stages, fairness constraints result in larger gender reemployment gaps.



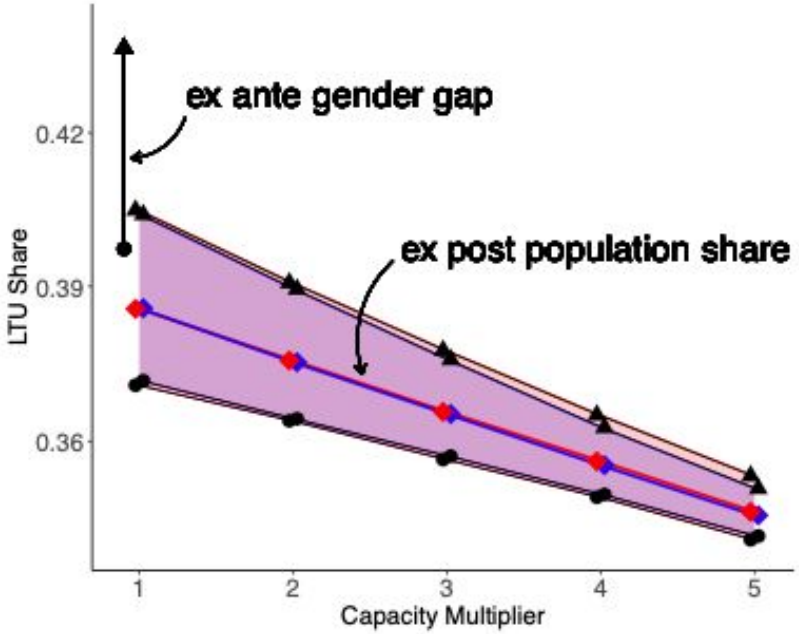
(c) Austrian Prioritization and Optimal Program



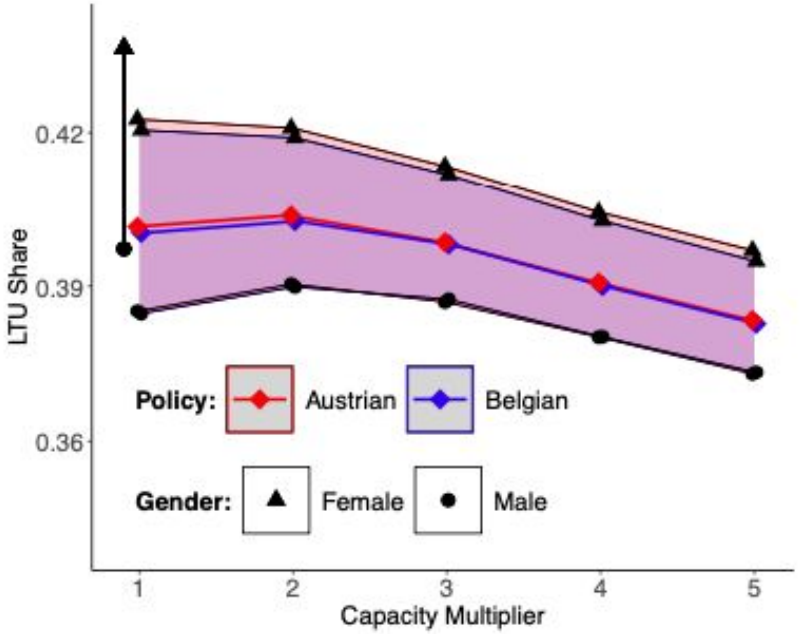
(d) Austrian Prioritization and Random Program

Hawks and Doves

We find no efficiency gains in neglecting those at highest risk: Austrian prioritization is slightly worse overall and creates larger gender gaps.



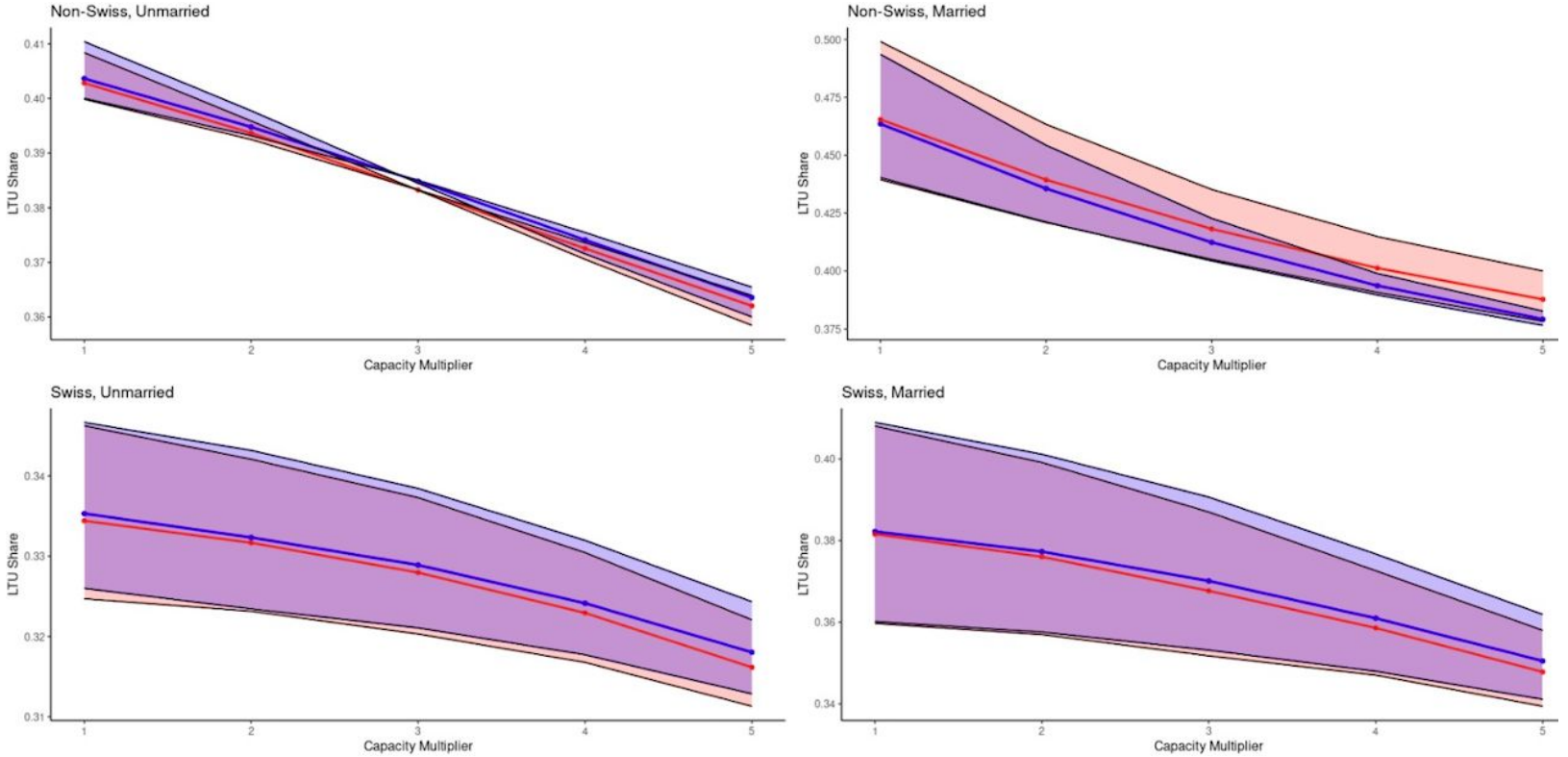
(a) Optimal Program



(b) Random Program

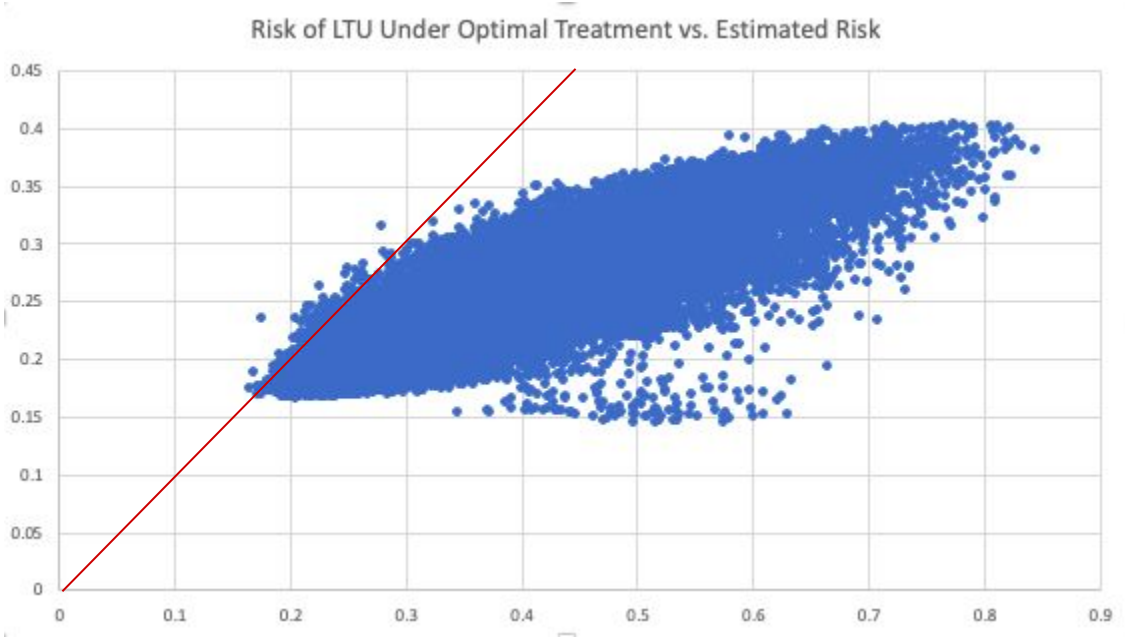
Hawks and Doves

Belgian policy narrows gender gaps among the least advantaged (married, non-citizens).



Actual and Counterfactual Risk

The riskiest people have only a 40% risk of becoming long-term unemployed under **optimal** treatment.



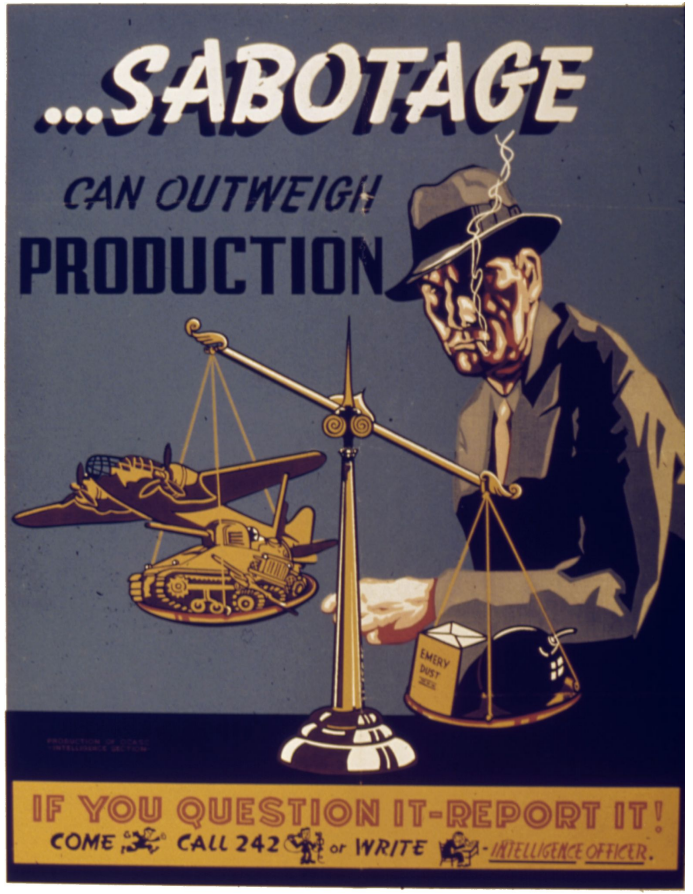
Takeaways

The Fair Distribution of Predictions, or Social Goods?

‘Fair’ risk scores are self-sabotaging!

By making the risk of men and women look similar, fairness constraints give women fewer spots in effective programs than they would get otherwise.

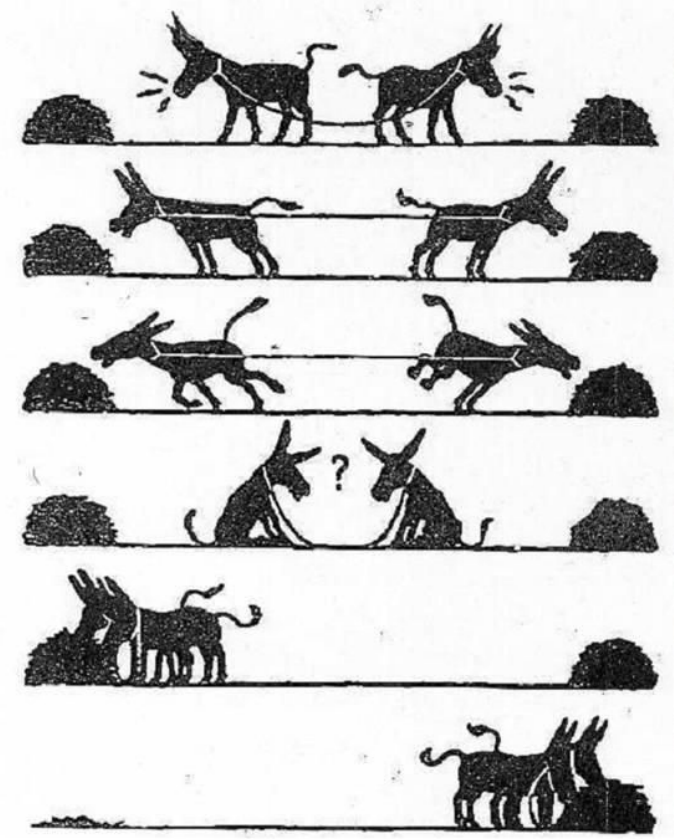
Constraining risk scores to be ‘fair’ undermines policy aimed at lowering overall unemployment and narrowing gender gaps.



Philosophy of Science Vindicated

There is no trade-off between accurate (counterfactual) predictions and the fair distribution of social goods.

Better individualized estimates of program effectiveness yield better overall rates of long-term unemployment **and** smaller gender reemployment gaps, regardless of other choices.



Dove Supremacy

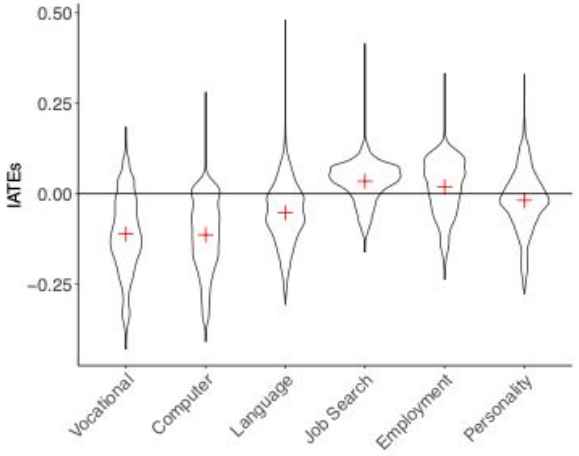
Policies that deny resources to the highest risk are no more efficient than those that target the highest risk.

Someone is only ‘hopeless’ if they wouldn’t do well under the **right** treatment.

DALL-E: A propaganda poster advocating for absolute dove supremacy in the struggle against hawks.

Shortcomings: Counterfactual Prediction

We rely on estimates of counterfactual outcomes under interventions.



(a) Individualized Average Treatment Effects.

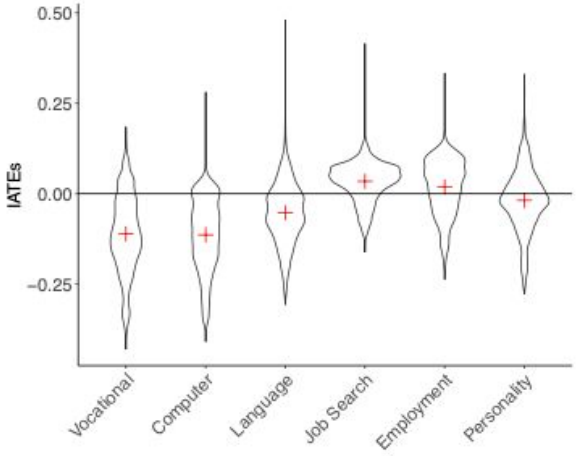
	ATE	SE	95%-CI
Vocational	-11.12	0.06	[-11.12, -11.12]
Computer	-11.37	0.05	[-11.37, -11.37]
Language	-5.25	0.04	[-5.26, -5.25]
Job Search	3.43	0.03	[3.43, 3.43]
Employment	1.83	0.04	[1.83, 1.83]
Personality	-1.84	0.04	[-1.84, -1.84]

(b) Average Treatment Effects in percentage points, standard errors, and 95% confidence intervals. Negative treatment effects imply a lower risk of becoming long-term unemployed.

Fig. 3. (Individualized) average treatment effects for the six labor market programs. No program serves as baseline.

Shortcomings: Counterfactual Prediction

Errors are hard to control, and their effects on the outcomes of allocation are difficult to predict: inductive risk is relatively unmanaged.



(a) Individualized Average Treatment Effects.

	ATE	SE	95%-CI
Vocational	-11.12	0.06	[-11.12, -11.12]
Computer	-11.37	0.05	[-11.37, -11.37]
Language	-5.25	0.04	[-5.26, -5.25]
Job Search	3.43	0.03	[3.43, 3.43]
Employment	1.83	0.04	[1.83, 1.83]
Personality	-1.84	0.04	[-1.84, -1.84]

(b) Average Treatment Effects in percentage points, standard errors, and 95% confidence intervals. Negative treatment effects imply a lower risk of becoming long-term unemployed.

Fig. 3. (Individualized) average treatment effects for the six labor market programs. No program serves as baseline.

Prediction Focused Learning

Prediction-focused learning factors policy applications into

(1) a **prediction step**, in which you try to predict outcomes as “accurately” as possible, and

(2) an **allocation step**, in which predictions from step (1) are fed into an optimization procedure that outputs the most socially desirable allocation of resources.

Prediction Focused Learning

Prediction-focused learning factors policy applications into

(1) a **prediction step**, in which you try to predict outcomes as “accurately” as possible, and

(2) an **allocation step**, in which predictions from step (1) are fed into an optimization procedure that outputs the most socially desirable allocation of resources.

While *perfect* predictions would lead to optimal allocations, it is sometimes more practical to estimate allocation policies “directly”.

Decision Focused Learning

Decision-focused learning favors an “end-to-end” system in which steps (1-2) are repeated until an optimal policy is found

(1) **predictions** are used to arrive at optimal allocations and

(2) errors in allocation are back-propagated to update predictions in part (1).

Decision-Focused Learning: Foundations, State of the Art, Benchmark and Future Opportunities

Jayanta Mandi * JAYANTA.MANDI@KULEUVEN.BE
KU Leuven, Belgium

James Kotary † JK4PN@VIRGINIA.EDU
University of Virginia, USA

Senne Berden SENNE.BERDEN@KULEUVEN.BE
KU Leuven, Belgium

Maxime Mulamba MAXIME.MULAMBA@VUB.BE
Vrije Universiteit Brussel, Belgium

Víctor Bucarey VICTOR.BUCAREY@UOH.CL
Universidad de O'Higgins, Chile

Tias Guns TIAS.GUNS@KULEUVEN.BE
KU Leuven, Belgium

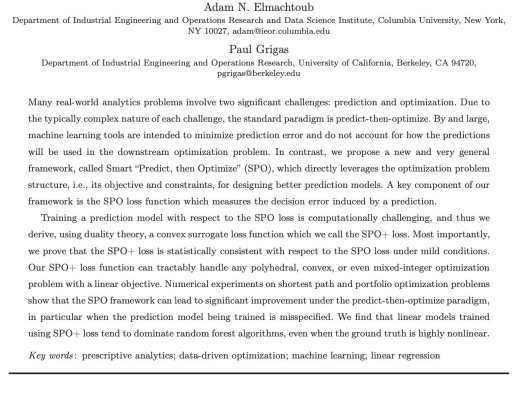
Ferdinando Fioretto FIORETTO@VIRGINIA.EDU
University of Virginia, USA

Abstract

Decision-focused learning (DFL) is an emerging paradigm that integrates machine learning (ML) and constrained optimization to enhance decision quality by training ML models in an end-to-end system. This approach shows significant potential to revolutionize combinatorial decision-making in real-world applications that operate under uncertainty, where estimating unknown parameters within decision models is a major challenge. This paper presents a comprehensive review of DFL, providing an in-depth analysis of both gradient-based and gradient-free techniques used to combine ML and constrained optimization. It evaluates the strengths and limitations of these techniques and includes an extensive empirical evaluation of eleven methods across seven problems. The survey also offers insights into recent advancements and future research directions in DFL.

Code and benchmark: `https://github.com/PredOpt/predopt-benchmarks`

1. Introduction



Smart “Predict, then Optimize”

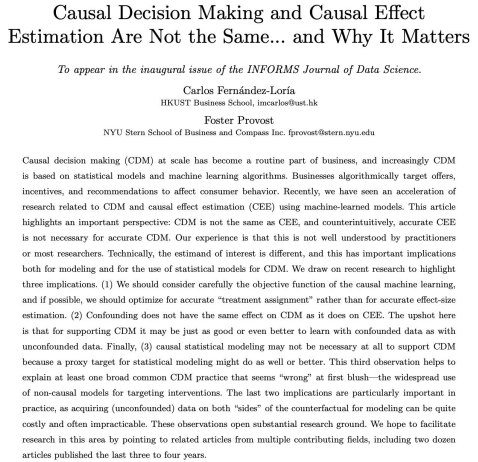
Adam N. Elmachtoub
Department of Industrial Engineering and Operations Research and Data Science Institute, Columbia University, New York, NY 10027, adam@ieor.columbia.edu

Paul Grigas
Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720, pgrigas@berkeley.edu

Many real-world analytics problems involve two significant challenges: prediction and optimization. Due to the typically complex nature of each challenge, the standard paradigm is predict-then-optimize. By and large, machine learning tools are intended to minimize prediction error and do not account for how the predictions will be used in the downstream optimization problem. In contrast, we propose a new and very general framework, called Smart “Predict, then Optimize” (SPO), which directly leverages the optimization problem structure, i.e., its objective and constraints, for designing better prediction models. A key component of our framework is the SPO loss function which measures the decision error induced by a prediction.

Training a prediction model with respect to the SPO loss is computationally challenging, and thus we derive, using duality theory, a convex surrogate loss function which we call the SPO+ loss. Most importantly, we prove that the SPO+ loss is statistically consistent with respect to the SPO loss under mild conditions. Our SPO+ loss function can tractably handle any polyhedral, convex, or even mixed-integer optimization problem with a linear objective. Numerical experiments on shortest path and portfolio optimization problems show that the SPO framework can lead to significant improvement under the predict-then-optimize paradigm, in particular when the prediction model being trained is misspecified. We find that linear models trained using SPO+ loss tend to dominate random forest algorithms, even when the ground truth is highly nonlinear.

Key words: prescriptive analytics; data-driven optimization; machine learning; linear regression



Causal Decision Making and Causal Effect Estimation Are Not the Same... and Why It Matters

To appear in the inaugural issue of the INFORMS Journal of Data Science.

Carlos Fernández-Loría
HKUST Business School, imcarlos@ust.hk

Foster Provost
NYU Stern School of Business and Compass Inc. fprovost@stern.nyu.edu

Causal decision making (CDM) at scale has become a routine part of business, and increasingly CDM is based on statistical models and machine learning algorithms. Businesses algorithmically target offers, incentives, and recommendations to affect consumer behavior. Recently, we have seen an acceleration of research related to CDM and causal effect estimation (CEE) using machine-learned models. This article highlights an important perspective: CDM is not the same as CEE, and counterintuitively, accurate CEE is not necessary for accurate CDM. Our experience is that this is not well understood by practitioners or most researchers. Technically, the estimand of interest is different, and this has important implications both for modeling and for the use of statistical models for CDM. We draw on recent research to highlight three implications. (1) We should consider carefully the objective function of the causal machine learning, and if possible, we should optimize for accurate “treatment assignment” rather than for accurate effect-size estimation. (2) Confounding does not have the same effect on CDM as it does on CEE. The upshot here is that for supporting CDM it may be just as good or even better to learn with confounded data as with unconfounded data. Finally, (3) causal statistical modeling may not be necessary at all to support CDM because a proxy target for statistical modeling might do as well or better. This third observation helps to explain at least one broad common CDM practice that seems “wrong” at first blush—the widespread use of non-causal models for targeting interventions. The last two implications are particularly important in practice, as acquiring (unconfounded) data on both “sides” of the counterfactual for modeling can be quite costly and often impracticable. These observations open substantial research ground. We hope to facilitate research in this area by pointing to related articles from multiple contributing fields, including two dozen articles published the last three to four years.

Decision Focused Learning

Decision-focused learning favors an “end-to-end” system in which steps (1-2) are repeated until an optimal policy is found

(1) **predictions** are used to arrive at optimal allocations and

(2) errors in allocation are back-propagated to update predictions in part (1).

Decision-focused learners argue that (1) there is no independent predictive context in which only the value of predictive accuracy reigns and (2) the only relevant prediction errors are the ones that affect allocation decisions. Optimize for **optimal allocation**, rather than accurate estimation of treatment effects.

Decision Focused Learning

Decision-focused learning favors an "end-to-end" system in which steps (1-2) are repeated until an optimal policy is found

(1) **predictions** are used to arrive at optimal allocations and

(2) errors in allocation are back-propagated to update predictions in part (1).

Decision-focused learners argue that (1) there is no independent predictive context in which only the value of predictive accuracy reigns and (2) the only relevant prediction errors are the ones that affect allocation decisions. Optimize for **optimal allocation**, rather than accurate estimation of treatment effects.

Pragmatic encroachment in ML! Managing inductive risks is more important than predictive "accuracy."

Decision Focused Learning: Difficulties

DFL faces many difficulties:

(1) Is it even possible to back-propagate errors in policy settings? Where do you get reliable signals of policy error?

(2) Predictions are finely tuned to a policy goal and cannot be re-used in other policy contexts.

(3) ML engineers are empowered to make normative decisions about appropriate (surrogate) local justice principles. Backprop favors *differentiable* loss functions.

(4) Cannot be used to adjudicate between different policies!

Thank You!

(1) Counterfactual Prediction

For each individual we predict their (counterfactual!) outcomes under each of the seven treatments.*

Identification assumes *Unconfoundedness*, *Common Support*, and *Stable Unit Treatment Value*.

Assumption 2.1.

(a) *Unconfoundedness:* $Y_i(w) \perp\!\!\!\perp W_i \mid X_i = x, \forall w \in \mathcal{W}, and x \in \mathcal{X}.$

(b) *Common support:* $0 < P[W_i = w \mid X_i = x] \equiv e_w(x), \forall w \in \mathcal{W} and x \in \mathcal{X}.$

(c) *Stable Unit Treatment Value Assumption (SUTVA):* $Y_i = Y_i(W_i).$

*This step is the most technical, using a relatively new technique called double-robust machine learning (Chernozhukov et al., 2018 and Knaus, 2022).

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal.

Michael C. Knaus. 2022. Double machine learning-based programme evaluation under unconfoundedness

Identifiability

Theorem (Identification of $P_{\text{post}}(Y = y \mid A = a)$)

Suppose that CONSISTENCY, UNCONFOUNDEDNESS, NO UNPRECEDENTED DECISIONS, STABLE CATE and NO FEEDBACK hold. Suppose also that $P_{\text{post}}(A = a) > 0$. Then, $P_{\text{post}}(Y = y \mid A = a)$ is given by

$$\sum_{(x,d) \in \Pi_{\text{post}}} P_{\text{pre}}(Y = y \mid A = a, X = x, D = d) P_{\text{pre}}(X = x \mid A = a) \underbrace{P_{\text{post}}(D = d \mid A = a, X = x)}_{\text{ALGORITHMIC EFFECT}},$$

where $\Pi_t = \{(x, d) \in \mathcal{X} \times \mathcal{D} : P_t(X = x, D = d \mid A = a) > 0\}$.

Sebastian Zezulka, Konstantin Genin (2023). *Performativity and Prospective Fairness.* Accepted: Algorithmic Fairness Through the Lens of Time, NeurIPS Workshop.

Identifiability

$$Y = \sum_{d \in \mathcal{D}} Y^d \mathbb{1}[D = d]. \qquad \text{CONSISTENCY}$$

$$Y^d \perp\!\!\!\perp_t D \mid A, X. \qquad \text{UNCONFOUNDEDNESS}$$

$$P_{\text{pre}}(D = d \mid A = a, X = x) > 0 \text{ if } P_{\text{post}}(D = d \mid A = a, X = x) > 0 \qquad \text{NO UNPRECEDENTED DECISIONS}$$

$$P_{\text{pre}}\left(Y^d \mid A = a, X = x\right) = P_{\text{post}}\left(Y^d \mid A = a, X = x\right) \qquad \text{STABLE CATE}$$

$$P_{\text{pre}}\left(A = a, X = x\right) = P_{\text{post}}\left(A = a, X = x\right) \qquad \text{NO FEEDBACK}$$

- General *ceteris paribus* assumption: We want to isolate the effect of the policy.
- Assumptions do **not** rule out the intended direct effect of the algorithm on the decisions.

$$P_{\text{pre}}\left(D = d \mid A = a, X = x\right) \neq P_{\text{post}}\left(D = d \mid A = a, X = x\right) \qquad \text{ALGORITHMIC EFFECT}$$

Sebastian Zezulka, Konstantin Genin (2023). *Performativity and Prospective Fairness.* Accepted: Algorithmic Fairness Through the Lens of Time, NeurIPS Workshop.

Unconfoundedness

$$Y^d \perp\!\!\!\perp_t D \mid A, X. \qquad \text{UNCONFOUNDEDNESS}$$

Assumes that all the common causes of treatment and outcome are observed.

More plausible for rich administrative datasets.

Violated if caseworkers make their decisions on the basis of unrecorded properties.

Sebastian Zezulka, Konstantin Genin (2023). *Performativity and Prospective Fairness.* Accepted: Algorithmic Fairness Through the Lens of Time, NeurIPS Workshop.

No Unprecedented Decisions

$$P_{\text{pre}}(D = d \mid A = a, X = x) > 0 \text{ if } P_{\text{post}}(D = d \mid A = a, X = x) > 0 \qquad \text{No Unprecedented Decisions}$$

There are no genuinely **novel** combination of treatment and covariates (propensity scores are non-zero).

Sebastian Zezulka, Konstantin Genin (2023). *Performativity and Prospective Fairness.* Accepted: Algorithmic Fairness Through the Lens of Time, NeurIPS Workshop.

Stable CATE

$$P_{\text{pre}}\left(Y^d \mid A = a, X = x\right) = P_{\text{post}}\left(Y^d \mid A = a, X = x\right) \qquad \text{STABLE CATE}$$

Assumes that the effectiveness of the programs (for people with $A = a, X = x$) does not change, so long as all that has changed is the way we *allocate* people to programs.

This assumption could be violated if e.g., a program works primarily by making some better off only at the expense of others—if everyone were to receive such a program, it would have no effect.

Sebastian Zezulka, Konstantin Genin (2023). *Performativity and Prospective Fairness.* Accepted: Algorithmic Fairness Through the Lens of Time, NeurIPS Workshop.

No Feedback

$$P_{\text{pre}}(A = a, X = x) = P_{\text{post}}(A = a, X = x) \qquad \text{NO FEEDBACK}$$

No Feedback assumes that the baseline covariates of the recently employed are identically distributed pre- and post-deployment. Strictly speaking, this is false, since the decisions of caseworkers will affect the covariates of those who re-enter employment and some of them will, eventually, become unemployed again.

But since the pool of employed is much larger than the pool of unemployed, the policies of the employment service have much larger effects on the latter than the former. For this reason, we may hope that feedback effects are not **too** significant

Sebastian Zezulka, Konstantin Genin (2023). *Performativity and Prospective Fairness.* Accepted: Algorithmic Fairness Through the Lens of Time, NeurIPS Workshop.

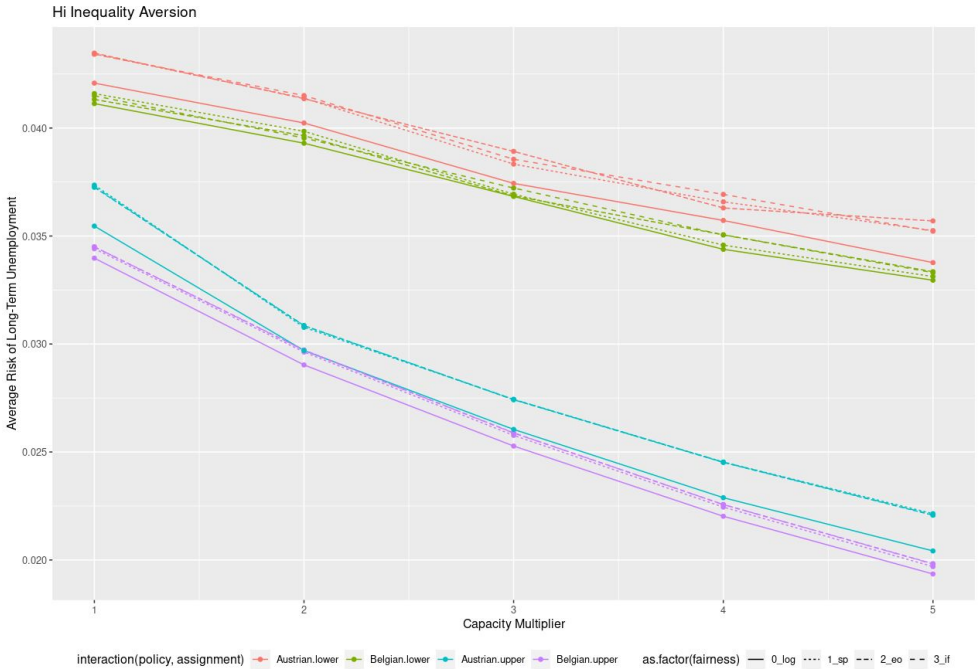
Welfare Analysis

To get a sense of the overall effects of the different policies, we looks at the average resulting (inequality-weighted) risk for each policy:

$$\sum_i E[Y_i]^{\frac{1}{1-\epsilon}},$$

Where ε = 0 (no); ε = .25 (low); ε = .5 (medium), or ε = .75 (high inequality aversion).

Welfare Analysis



Not only is the Austrian policy **less** efficient, but constraining the risk predictor to be ‘fair’ makes the (inequality-weighted) average outcome strictly worse!