Prediction, Performativity, and Potential Outcomes: Communicative Rationality in Prediction-Allocation Problems

Sebastian Zezulka SEBASTIAN.ZEZULKA@UNI-TUEBINGEN.DE
University of Tübingen
Tübingen, Germany

Konstantin Genin U6068641@UTAH.EDU
University of Utah
Salt Lake City, Utah, USA

Abstract

In prediction-allocation problems, predictions are used to allocate social goods. Standard methodology recommends factoring these problems into two stages: first, individual outcomes are predicted as accurately as possible; second, allocations are made based on these predictions. But when predictions inform allocation, they causally influence outcomes. The "performative" nature of these predictions entangles epistemic and pragmatic considerations, making notions of accuracy ambiguous. Two families of responses propose to correct the standard procedure. Emphasizing the epistemic, *endogenization* recommends making predictions that tend to make themselves true. *Pragmatic encroachment* recommends making the predictions with the best distributive consequences. We argue that both responses are misguided. The former undermines the normative goals of allocation. The latter, despite being instrumentally rational, has a distorting effect on communication. When predictions are expressions of substantive normative commitments, they cannot help resolve normative disputes among policymakers and street-level bureaucrats. That is, they are not communicatively rational. We argue that one does not have to trade off instrumental and communicative rationality and introduce two criteria of adequacy for predictions: they should be *decision*- and *discourse*-supportive. These criteria can be met by predicting counterfactual outcomes. We argue that predictions satisfying these criteria can serve as public reasons.

Keywords: Algorithmic Decision-Making, Performativity, Counterfactual Prediction, Communicative Rationality, Instrumental Rationality

1 Introduction

In public administration, statistical predictions routinely mediate the allocation of social goods: we call such exercises *prediction-allocation problems.* The choice of predictive methodology partly determines whether allocations are distributively just. But it also determines whether discourse, among street-level bureaucrats and policymakers alike, can answer to the demands of public reason (Rawls, 1997) and communicative rationality (Habermas, 1984). In this paper, our fundamental concern is how methodological choices in statistical modeling either enable or foreclose meaningful deliberation. We argue that predictions used to inform the allocation of social goods must be able to serve as *public reasons* i.e., as reasons that command broad acceptance as plausible bases for action and deliberation (Quong,

Presented as *non-archival* paper at Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '25), *November 5–7, 2025, Pittsburgh, PA, USA.* The abstract can be found at
https://doi.org/10.1145/3757887.3767685

2022; Binns, 2017). Requiring predictions to answer to high standards of publicity can be understood as a plaidoyer for the value-free ideal. However, whether this ideal supports or distorts deliberation depends delicately on whether we choose to predict actual or counterfactual outcomes. In typical prediction-allocation problems, counterfactual outcomes are the appropriate targets of estimation. Moreover, prominent strategies for bringing non-epistemic values to bear on predictions of actual (or counterfactual) outcomes inevitably have distorting effects on discourse.

Prediction-allocation problems are inherently political. Policymakers must set the goals of allocation and decide what kind of resources are to be made available to whom and in what quantities. If resources are scarce, they must consider principles to allocate them.[1] Street-level bureaucrats must decide who gets desirable social goods and how to balance the normative demands of providing optimal care for the individual, treating all clients fairly, and enforcing the rules of their institution (Lipsky, 2010; Zacka, 2017). Statistical predictions commonly inform the design and implementation of such mechanisms of allocation. Predictions of the risk of dropping out of school inform the allocation of prevention programs (Perdomo et al., 2023); predictions of the risk of child abuse inform the allocation of investigators (Coston et al., 2020); predictions of mortality risk inform the allocation of healthcare resources (Caruana et al., 2015). Throughout the paper, we illustrate prediction-allocation problems with two case studies: Unemployment, where caseworkers decide who among the recently unemployed should receive job training and Pneumonia, where doctors decide whom to admit to the hospital. In these settings, predicted outcomes—risk of mortality and long-term unemployment—are *intended* to inform the choice of intervention.

Received statistical wisdom recommends factoring prediction-allocation problems into two stages.[2] In the first stage, demographic, individual, and environmental data (X) is used to arrive at predictions ($\hat{Y}$) of the outcome of interest (Y) that are as accurate as possible.[3] In the second stage, individual predictions are used to arrive at allocation decisions (D) that are as optimal as possible in light of the normative goals of allocation. Importantly, the consequences of the decisions are themselves causally relevant to the realized individual outcomes (Y). In cases like these, statistical predictions can have a variety of undesired effects. In Unemployment, seemingly accurate predictions of high risk can lead to self-fulfilling prophecies. In Pneumonia, seemingly accurate predictions of low risk can mislead triage decisions, leading to self-defeating prophecies. These effects are typically blamed on *performativity*, the capacity of predictions to causally influence the outcomes they are meant to foresee. Performativity entangles pragmatic and epistemic considerations, making notions of accurate prediction ambiguous.

The standard response to the problem of performativity is to propose strategies for *managing* performative effects. We distinguish two families of such proposals. The first recommends *endogenizing* performative effects and, in this way, realizing the narrowly defined goal of predictive accuracy. Endogenization often leads to normatively undesirable outcomes.

[1] Allocation principles recommend that decisions respect factors like desert, need, efficiency, or equality and are, in turn, operationalizations of normative commitments like prioritarianism or utlitarianism (Elster, 1992; Kuppler et al., 2022).

[2] Debates in algorithmic fairness typically focus on defining statistical criteria of non-discrimination (Hedden, 2021; Eva, 2022; Beigang, 2023). Recently, there is increased emphasis on the distributive effects of predictions (Shirali et al., 2024; Zezulka and Genin, 2024).

[3] In these cases, accuracy is usually operationalized as mean squared-error or zero-one loss.

Moreover, it does not respect the nature of prediction-allocation problems: predictions are not mere forecasts that we must get right, but are intended to support decision-making. The second recommends *pragmatic encroachment:* the adequacy of predictions is no longer merely an epistemic question of accuracy, but assessed by the normative desirability of their likely consequences. These strategies include proposals for managing inductive risk from the philosophy of science, optimal steering proposals from computer science, and methodologies for decision-focused and policy-learning from econometrics and the policy sciences.

When we adopt a pragmatic encroachment strategy, predictions become expressions of substantive normative commitments. The resulting predictions can support the implementation of a pre-specified allocation principle, avoiding undesirable performative effects. But they cannot help to resolve normative disputes among policymakers and street-level bureaucrats who do not already agree on a principle of allocation. That is to say, pragmatic encroachment strategies are *practically*, but not *communicatively*, rational. Although often assumed away in the methodological literature, the political goals of allocation are typically not agreed upon before statistical methods are applied—these goals are the outcomes of complex policy discussions. And even if a policy decision has been made, street-level bureaucrats still face individual decisions not fully determined by high-level policy. Standard statistical predictions, as well as proposals for managing their performative consequences, all have distorting effects on policy discussions and street-level decision-making.

We argue that the difficulties encountered in prediction-allocation problems have been misdiagnosed. The problem in cases like Unemployment and Pneumonia is not the performative nature of predictions as such, but the *infelicity* of standard statistical predictions in performative contexts. Standard predictions *reify* social outcomes: statistical regularities are presented as natural laws to which we can accommodate ourselves, but which we cannot change. That obscures the fact that outcomes are, to a significant degree, influenced by our decisions.

We argue that one does not have to trade the epistemic against the pragmatic as the received diagnosis suggests. Discussions among policymakers and street-level bureaucrats can be productively informed by *counterfactual* predictions, if these are accurately estimated, clearly communicated and carefully interpreted in light of inevitable uncertainties. Accurate counterfactual, rather than actual, predictions answer the demands of both practical and communicative rationality. Specifically, we introduce two criteria of adequacy for predictions: they must be decision- and discourse-supportive. Predictions are *decision-supportive* if they allow an agent to effectively evaluate policy options in light of her own goals and values. Predictions are *discourse-supportive* if they allow a number of discourse participants, with potentially conflicting normative commitments, to evaluate policy options in light of the goals and values of all participants in the discourse. In other words: when they can serve as public reasons.

This paper is structured as follows. In section 2, we introduce our two case studies and the problem of self-fulfilling and self-defeating prophecies. We discuss the received etiology for these problems in section 3 and argue that the two main prescriptions for managing performativity, endogenization (4) and pragmatic encroachment (5), are inadequate. In section 6, we present our alternative diagnosis: standard statistical predictions give rise to perverse effects because they are pragmatically ill-formed in prediction-allocation contexts;

accurate counterfactual predictions do not suffer from these defects. In section 7, we argue that accurate counterfactual predictions are not only decision- but also discourse-supportive.

2 Prediction-Allocation Problems

A proliferation of examples in which predictions, although accurate by the usual statistical standards, fail to support the normative goals of allocation, puts the wisdom of the standard two-stage approach to prediction-allocation problems into question. We give two stylized examples, which will guide the subsequent discussion.

> Unemployment. An algorithm is developed to predict the risk of long-term unemployment for recently unemployed workers reporting to a public employment service. At the street level, caseworkers use these scores to decide whether to enroll workers in one of several training programs. Space in the courses is limited, and some programs are known to be, on average, more effective than others. Meanwhile, a debate is underway among policymakers. Policy hawks, motivated by considerations of efficiency, recommend reserving spots in desirable programs for individuals of moderate risk and withholding resources from those at high and low risk. Policy doves, motivated by prioritarian convictions, recommend reserving spots in desirable programs for those at highest risk. Doves claim that a hawkish policy will only exacerbate structural inequalities, such as the longstanding gender gap in reemployment prospects. Hawks claim that a dovish policy will only waste resources on individuals who will anyway fail to benefit from them. At the street level, allocation decisions reflect the sympathies of individual caseworkers for one or another pole of this debate.

Based on a recent case in Austria (Allhutter et al., 2020; Achterhold et al., 2025), this example is a case in which social statistics are appealed to in a debate between competing normative goals. Using administrative data from Switzerland, Zezulka and Genin (2024) find that high-risk workers were historically less likely to be assigned to the most effective programs, even though they would stand to benefit the most. In other words, were the hawks to prevail, a good would be withheld from workers who are considered high risk partly because they were historically denied access to it. Moreover, their subsequent poor outcomes would be in part due to their smaller share of the good. This would be a case of self-fulfilling prophecy. Zezulka and Genin (2024) find that a dovish allocation is more effective at reducing the gender gap in reemployment prospects without incurring a higher average incidence of long-term unemployment. This is because training programs tend to be more effective for higher-risk workers. In this case, both hawks and doves can be satisfied: goods can be allocated in a prioritarian fashion without sacrificing overall efficiency.

Unemployment is a paradigmatic example in which the data used to train a prediction algorithm reflect historical injustices. It is natural to think that these injustices are responsible for the failure of predictions to support the normative goals of allocation. However, the next example shows that similar difficulties can arise even in the absence of historical injustice.

> Pneumonia.[4] An algorithm is developed to predict the mortality risk of pneumonia patients at hospital intake. The risk scores are intended to help physicians decide whether to admit the patient into hospital, or to follow up with outpatient treatment. After the modeling is complete, it is found that a history of asthma *lowers* the predicted risk of mortality. A surprising result, given that both pneumonia and asthma are pulmonary diseases. Eventually, the reason becomes clear: pneumonia patients with a history of asthma are lower risk because, historically, they are admitted directly into intensive care and, consequently, have better outcomes than the overall pneumonia population. However, were physicians to admit only high risk patients, this association would be reversed: asthmatics would be treated outpatient and have worse outcomes than the overall pneumonia population.

This example, based on a study from Pittsburgh (Cooper et al., 1997), is an instructive case of a potential self-defeating prophecy. If doctors admitted into hospital only those patients predicted to be at high risk, a good would be denied to asthmatics because, perversely, they have received this good historically. Moreover, their subsequent poor outcomes would be in part due to their smaller share of the good. Note that this could happen even though patient care is of high quality, predictions are accurate by the usual statistical standards and everyone involved is behaving responsibly.

With these case studies, we have identified three shortcomings of the two-stage approach to prediction-allocation problems. First, as in Pneumonia, seemingly accurate predictions can mislead decision-makers and thereby undermine the very policy goal they are intended to support. Second, as in Unemployment, seemingly accurate predictions can give rise to a kind of statistical fatalism where historical outcomes, themselves the product of earlier policy choices, are misrepresented as inevitable (Hacking, 1990, 115). Third, accurate predictions fail to provide the evidential basis for prospective policy evaluation and, *a fortiori*, for deliberation over the normative goals of allocation. In Unemployment, accurate predictions of actual outcomes do not reveal that the policy dispute between hawks and doves is in fact a pseudo-conflict, since both can satisfy their respective goals without frustrating the goals of the other.

Note that self-defeating and self-fulfilling prophecies are not necessarily undesirable. Usually, we would like predictions of high risk to be self-defeating and predictions of low risk to be self-fulfilling but not, as in the previous examples, vice-versa. From an epistemic perspective, self-defeating prophecies of low risk are more benign than self-fulfilling prophecies of high risk. Although the human costs in Pneumonia would be significant, physicians and modelers would at least be alerted to the fact that something had gone wrong when their predictions of low mortality for asthmatics are not borne out. Self-fulfilling prophecies of high-risk, as in Unemployment, are more insidious, since modelers and policy-makers could mistake the fact that their predictions are borne out for evidence that things could not have been different.

[4] We are indebted to Tal (2023) for this example.

3 Performativity

A (statistical) model is called *performative* when, in addition to serving epistemic purposes, it causally influences the system it is meant to represent. Performativity is ubiquitous in the social sciences and operates on many levels.[5] 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, 1995). Individuals may react to predictions made by an epidemiological model of viral spread by limiting social interaction, thereby vitiating the assumptions of the model (Basshuysen et al., 2021; Winsberg and Harvard, 2022). Individuals may react to a credit-scoring system by strategically manipulating the way they are represented in data e.g., by opening new lines of credit (Dalvi et al., 2004; Hu et al., 2019). These are instances of *macro*-performativity, in which models have diffuse effects on entire populations, either by entering themselves into public discourse, making something common knowledge, making aggregate predictions, or restructuring prevailing incentives. These effects are often unintended and typically difficult to anticipate with any precision. For our purposes, we are interested in *micro*-performativity, in which a local prediction about an individual causally influences their individual future outcome.[6] These effects are partially intended and somewhat easier to anticipate (Zezulka and Genin, 2023).

When informing allocation decisions, predictions do not only play a descriptive role: they also have a causal effect on the outcome. In other words, the performative effects are *intended.* The purpose of predicting mortality in Pneumonia is to inform triage decisions and, hopefully, lower mortality rates; the purpose of predicting the risk of long-term unemployment is to inform assignment to training programs and, hopefully, lower rates of long-term unemployment (and decrease gender gaps). Individual predictions change the likelihood of receiving a social good and, thereby, the distribution of outcomes. In statistical jargon: they change the propensity of treatment.

Performativity drastically complicates the evaluation of predictions, as the standard by which to evaluate them is no longer clear. For example, the commonly used ”label matching conception of accuracy” (Tal, 2023) for classification tasks, the probability that the prediction $\hat{y}$ equals the observed outcome y, is uninformative under performativity. For self-defeating prophecies of low risk, as in Pneumonia, label matching accuracy on a test set can be high, not indicating the potential for drastic failure once the model is deployed. For self-fulfilling prophecies of high risk, label-matching accuracy will remain high after deployment, but for the wrong reason. If policy hawks prevail in Unemployment, individuals predicted to be at high risk would be denied access to resources, trapping them in a cycle of accurate (because self-fulfilling) predictions. More generally, performative effects challenge the standard paradigm in which a statistical model is judged by its ability to *generalize.* A model is said to generalize if its predictions are expected to be accurate on future instances drawn from the *same* distribution as the training samples (von Luxburg

[5] For Austin (1962), performative speech acts are those that not only describe but also establish states of affairs. In machine learning, performative predictions are those that induce distribution shifts (Perdomo et al., 2020). There is longstanding debate on the role of performativity in the social sciences (Merton, 1948; MacKenzie, 2006; Kopec, 2011; Guala, 2016). A similar debate orients itself around reflexivity (Buck, 1963; Romanos, 1973; Northcott, 2022).

[6] Mendler-Dünner et al. (2022) call this “outcome performativity”.

and Schölkopf, 2011; Grote et al., 2024). In prediction-allocation problems, we expect the deployment of predictions to *change* the distribution of future outcomes.

In the following, we distinguish two families of responses to performativity. The first, *endogenization*, recommends evaluating performative predictions solely based on accuracy in the distribution of outcomes that would be induced by their deployment. The second, *pragmatic encroachment*, recommends evaluating performative predictions based, at least in part, on the normative desirability of the distribution of outcomes they induce. We discuss both in turn.

4 Endogenizing Performativity

A first response to the problem of performativity is to treat it solely as an *epistemic* problem. The proposed strategy is to anticipate performative effects and to make the prediction $\hat{Y} = y$ which, taking account of its effect on future outcomes, brings about the outcome $Y = y$. Thus, the modeler proactively takes into account the effect of her own predictions on the system she is modeling. Khosrowi (2023) calls this *endogenizing* performativity. In a slogan: the modeler makes prophecies which tend to fulfill themselves. If successful, the predictions would be accurate in the "usual" epistemic sense.

Endogenization strategies draw on the distinction between private and public predictions. Private predictions are those that, in principle, cannot influence outcomes, perhaps because they are hidden in a desk drawer. In contrast, public predictions e.g., predictions of market prices or election results published in prominent venues, can change expectations, actions and, consequently, the outcomes which the predictions were meant to foresee. The challenge then is to integrate these responses into public predictions in order to find a point at which the predictions and outcomes coincide. Indeed, it can be shown that such points exists for a number of prediction problems, in the face of both macro- (Grunberg and Modigliani, 1954; Simon, 1954) and micro-performativity (Perdomo, 2025).

Endogenizing performative effects is inadequate in prediction-allocation problems for two reasons. First, based on a narrowly epistemic perspective, it problematizes self-defeating predictions, but lacks the conceptual resources to recognize self-fulfilling prophecies as a potential problem. Second, the goal of endogenization is to privatize public predictions: to make public predictions "as if" they are private predictions to which the system is indifferent. As we shall see, this is achieved only by disregarding the purposes for which predictions are intended.

Recall the setting of Pneumonia. If we are passive observers of the hospital, amusing ourselves by making bets on the outcomes of patients, our mortality predictions are "private" and do not affect treatment decisions or patient outcomes. Now imagine that these bets are made known to the doctors who, convinced of our cleverness, begin using them to inform their admissions decisions: our bets become public predictions. But now more patients that we had bet would survive (because they were asthmatics who previously received intensive treatment) are sent home, exposing them to high risk. Our amusement is disturbed: we had not taken into account the effect of our predictions on intake decisions.

'No fair!' we might cry. 'That was not part of the game.' Determined to have our fun, we try to endogenize the effects of our predictions by anticipating how the doctors would react to our bets. Holding fixed the intake policy, we know that increasing the predicted mortality leads to more intense treatment. So we must predict that asthmatics are at higher risk. But there is a limit to how high a risk we can assign them. For predicting too high mortality leads to more intensive treatment, resulting, on average, in good outcomes. Thus our predictions would once again be frustrated. It seems the only way to endogenize the effects of our bets is to assign asthmatics a middling risk, hoping that this confuses the doctors enough to result in middling outcomes. Unfortunately, this results in worse outcomes for asthmatics, frustrating the normative goals of allocation. Presenting itself as an epistemic response to the problem of performativity, endogenization has dire non-epistemic consequences. Therefore, endogenization is, implicitly, a rather dubious normative commitment (Khosrowi, 2023).

Endogenizing performative effects may well reach an optimal point in the narrow sense of allowing for accurate (public) predictions. But this is not even epistemically adequate, because it fails to respect the action-guiding purpose of predictions in prediction-allocation problems. In attempting to recover an *external* perspective on the system, endogenization neglects the *internal* purpose which predictions are meant to serve. The challenge is not to get any kind of accurate predictions but accurate predictions that respect the allocation problem that clinicians must solve. Predictions that are not supportive of the decision problem at hand cannot be considered good, even from an epistemic perspective.

In summary, endogenization cannot problematize self-fulfilling prophecies and fails to respect the epistemic structure of prediction-allocation problems. Operating under the rhetoric of the merely epistemic, it commits to realizing worse than achievable outcomes. These problems motivate a set of approaches we collect under the heading of "pragmatic encroachment" which make explicit the normative dimension of prediction-allocation problems.

5 Pragmatic encroachment

In epistemology, the pragmatic is said to encroach on the epistemic when practical considerations, such as the severity of the consequences of foreseeable errors, partly determine whether an agent *knows* a proposition (Stanley, 2005). We collect under its heading a number of proposals suggesting that practical considerations, such as the consequences of error, or the likely distributive outcomes, should inform predictions in performative settings. We first turn to inductive risk arguments (5.1) and then to technical proposals for operationalizing pragmatic encroachment: optimal steering (5.2) and decision-focused and policy learning (5.3).

5.1 Inductive Risk

Since evidence rarely establishes or refutes a scientific hypothesis with certainty, the acceptance and rejection of hypotheses is attended by *inductive risk*: the risk that the hypothesis is accepted or rejected in error. Inductive risk arguments attempt to strike a balance between the epistemic and the pragmatic by proclaiming a two-fold creed: first, try to be as accurate as possible; second, try to foresee and manage the non-epistemic consequences of errors.[7] In two classic papers, Rudner (1953) and Hempel (1960) argue that, in the presence of inductive risk, accepting or rejecting hypotheses is not a purely epistemic matter—the practical consequences of accepting or rejecting in error must also be weighed. We can only arrive at adequate rules of acceptance by considering the *practical* consequences of the possible outcomes (Hempel, 1960, 56). Extending the arguments of Rudner and Hempel, Douglas (2000) argues that non-epistemic values should inform *all* scientific decisions (including many methodological choices) that have foreseeable practical consequences.

We can understand inductive risk arguments as proposals for managing performativity: the practical consequences of over- or underestimating the risk must be taken into account when making decision-relevant predictions. In Pneumonia, one must take into account the practical consequences of under-estimating the mortality risk of an asthmatic patient presenting with pneumonia; in Unemployment of over-estimating the risk of long-term unemployment for a recently unemployed woman whose care responsibilities prevent her from moving for work.

There are three problems with this approach. First, the consequences of over- or under-estimation depend essentially on the policy context. When the policy context is open, or when the policy debate itself depends in part on the results of statistical modeling, it becomes nearly impossible to anticipate the consequences of misestimation. If, in the context of Unemployment, we expect that policy hawks will prevail, we should be more worried about over-estimating the risk for women with care responsibilities, since they will then be deprived of valuable training. If we expect that doves will prevail, we should be more worried about under-estimating their risk. Considerations of inductive risk do not seem to issue useful advice in this situation. Jeffrey makes a similar objection to inductive risk arguments: since a scientific law will be relevant in a great diversity of situations, it is meaningless to speak of *the* cost of mistaken acceptance or rejection (1956, 242). Perhaps there is some way in which we could incorporate our uncertainty about the policy into our considerations of the practical consequences of over- and under-estimation. But in performative contexts, the trouble with the inductive risk proposal lies deeper than what Jeffrey had foreseen.

Given that policy debates hinge in part on the results of our modeling, guarding against the consequences of errors flirts with technocratic steering of public deliberation. Decisions about which consequences to guard against, e.g. by under- or overestimating the risk for specific groups, shapes the prospects of any policy proposal under debate. Douglas issues a stark warning against deliberately steering public deliberation: such a course would violate ideals of honesty central to basic science and the basic ideal of democracy “that an elite few should not subvert the will of the many” (2009, 80). Bright (2017), interpreting Du Bois (1898), similarly warns that allowing pragmatic encroachment into social-scientific research might ultimately discredit science in the eyes of the public, preventing it from informing democratic decision making. Perhaps we can remain true to democratic ideals by adopting a plurality of modeling approaches and being transparent about the values that inform them

[7] See Elliott and Richards (2017) for case studies of inductive risk. Debates about the value-free ideal and the role of (non-epistemic) values in science are, of course, much broader, encompassing arguments from underdetermination (Longino, 1995), the use of *thick* concepts (Putnam, 2002; Alexandrova, 2018; Thoma, 2023), and the role of science in a democratic society (Kitcher, 2001, 2011; Hilligardt, 2024).

(Contessa, 2021). Nevertheless, there remains a third difficulty that cannot be addressed with the usual resources of inductive risk.

All inductive risk arguments assume that it is the consequences of *errors* which we must guard against. For Hempel, there is no reason to guard against the consequences of correctly accepting or rejecting an hypothesis, since this is “what science aims to achieve” (1960, 56). Similarly, Douglas (2009) argues that “the scientist should not think about the potential consequences of making an *accurate* empirical claim and slant their advice accordingly” (81, emphasis added). In performative contexts, however, it is not sufficient to guard only against errors. We must also guard against the potential consequences of accurate—because self-fulfilling—predictions.[8]

In Pneumonia, considerations of the practical consequences of error seem to favor overestimating the mortality risk of asthmatics. However, if we endogenize the performative effect of our predictions by assigning middling risk to asthmatics, considerations of inductive risk would be silent, since our predictions would be accurate, even though patients would suffer. In Unemployment, we are worried that, if hawks prevail, predictions of high risk *will* be accurate *but for the wrong reason*: namely, due to their self-fulfilling effects. In this case, considerations of inductive risk would also be silent. Indeed, inductive risk considerations cannot rule out endogenizing strategies since, were these to be adopted, predictions would turn out to be accurate. And when there are no errors whose consequences need to be managed, values are meant to fall silent. Thus, although inductive risk arguments seem to take seriously the normative goals of allocation, they are powerless to forestall normatively undesirable, although self-fulfilling, prophesies. Our conclusion is that considerations of inductive risk are inadequate for dealing with the kind of difficulties raised in performative contexts.

5.2 Optimal Steering

Arguments from inductive risk address the consequences of erroneous (rather than accurate) predictions because they are anxious not to endorse wishful thinking. Hypotheses should not be accepted or rejected *merely* on the basis of their desirable or undesirable moral qualities. But in performative contexts, we may sometimes get what we want merely by “wishing”. If our predictions can directly induce desirable consequences, it seems rather cold-blooded to insist too much on the epistemic virtues. Feeling the pull of this argument, Miller et al. (2021) and Perdomo (2023) suggest we might evaluate performative predictions directly by the normative desirability of the outcomes induced by their deployment. Perdomo (2023) is representative:

> If predictions can shape the world around us [. . .] the goals of prediction may not be to just accurately forecast future outcomes, but also to actively steer them towards socially desirable targets (3).

Optimizing for accurate forecasting recovers the endogenization approach discussed in section 4. But if we optimize for socially desirable targets, our predictions will be guided solely by the normative goals of allocation. What argument could we mount against optimizing for socially desirable targets?

[8] Citing concerns about reflexivity, Douglas (2009, 21) restricts scope to the natural sciences.

There are two problems. First, if we take the approach at face value, it recommends an insincere attitude toward the street-level consumers of predictions: predictions are made to steer decision-makers in socially desirable directions, rather than help them to arrive autonomously at good decisions.[9] This turns what is meant to be a cooperation between modelers and decision makers into a strategic interaction and, once decision makers begin to suspect that they are being steered, may ultimately undermine the goals of optimal steering. Second, optimal steering presupposes that the normative goals of allocation are settled before the modeling exercise gets underway. Thus, modelers either become technocrats who impose their values without democratic accountability, or passive consumers of values imported from the social spheres authorized to produce them. The important possibility foreclosed by optimal steering is that predictions might helpfully inform discourse about which normative goals are feasible and which algorithmic means are conducive to the various goals that we might find attractive. In other words, by "baking in" ready-made values, optimal steering deprives itself of the resources to adjudicate normative disputes, both at the street and the policy level.

Let us first consider the problem at street-level. Suppose that, faced with the prediction task in Unemployment, modelers are concerned with narrowing the gender gap in reemployment prospects. From their data, modelers discern that the caseworkers in Graz are more hawkish than those in Vienna: In Vienna, they reserve seats in effective programs for high-risk workers whereas in Graz, they reserve them for workers of moderate risk.[10] Knowing this, the modelers proactively embrace the impacts of their predictions and steer the caseworkers by labeling more recently unemployed Grazerinnen as moderate risks. At first, this has the intended consequence of narrowing the gender gap. But over time, the caseworkers grow suspicious, as they notice that risk estimates have shifted from historical levels. The previously receptive attitude of the caseworkers becomes mistrustful and strategic. Suspecting that risk levels are underestimated, caseworkers begin reserving seats only for low risk workers, more likely to be men. Thus, the goals of the modelers are frustrated: although they have attempted to anticipate the performative effect of their predictions, they have actually *changed* the performative effect. Performativity now emerges at a higher level as a change in the way caseworkers *react* to predictions.

We can gain a novel perspective on this situation by distinguishing between the *illocutionary* force and the *perlocutionary* effect of a prediction. Roughly, the illocutionary force of an utterance is its communicative significance, whereas its perlocutionary effect is the causal effect it has on the hearer. For example, the illocutionary force of 'You are standing on my foot.' is a demand that you move, whereas its perlocutionary effects on the hearer might include embarrassment. In the context of Unemployment the illocutionary force of 'This worker is at high risk of long-term unemployment.' is unclear, partly because it is ambiguous between the hortative 'This worker is likely to become long-term unemployed unless you intervene.' and the fatalistic 'This worker is likely to become long-term unemployed no matter what you do.'

The perlocutionary effect of the prediction on the caseworker depends on how its illocutionary force is disambiguated. How it is disambiguated depends, in turn, on the 'moral

[9] We can be understood as arguing for a version of boosting (Hertwig and Grüne-Yanoff, 2017) rather than nudging (Thaler and Sunstein, 2008).

[10] This scenario is freely invented for the benefit of clarity.

disposition' of the caseworker. Following Zacka (2017) and Vredenburgh (2023) we understand a caseworker's moral disposition as the entire set of dispositions she uses to exercise her discretion, which includes "epistemic dispositions to attend to certain information, say, but also moral dispositions to weigh values in a certain way, or practical dispositions to take certain means to one's ends" (Vredenburgh, 2023, 10). If the caseworker is doveish, she might interpret the prediction in the hortative sense and thereby prioritize the worker as an urgent case. A hawk might take it in the fatalistic sense and neglect the worker as a hopeless case.

A prediction intended to steer is a *perlocution*, a speech act performed, not with its conventional communicative aim, but with a perlocutionary aim (Austin, 1962, 101). A steering prediction departs from the conventional communicative aim of informing the hearer in favor of aiming for a desired causal effect. Habermas (1984) argues that perlocutions tend to turn a communicative interaction into a strategic one:

> 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 [effects] will be attributed to the speaker as intended results, the latter finds it necessary to offer explanations and denials, and if need be, apologies [. . .]. Otherwise, he has to expect that the other participants will feel deceived and adopt a strategic attitude in turn, steering away from action oriented to reaching understanding (294).

Optimal steering attempts to manage performativity by "going perlocutionary" and attending only to the causal effect the prediction has on the hearer, rather than the communicative function that it is conventionally meant to serve: informing the hearer about the consequences of various courses of action. It can be expected that perlocutionary predictions eventually undermine their aims, once decision-makers begin to suspect that predictions are not made with the sincere goal of informing their decisions. We will argue that a better solution to the problem of performativity is to clear up the ambiguous illocutionary force of these predictions, rather than abandoning their communicative function.

At the policy level, predictions intended to steer toward socially desirable targets cannot be used to inform debates about what normative principles should govern allocation. In such a debate, the socially desirable target is precisely what is at issue and cannot be assumed *ex ante*. If the modelers are not up-front about the steering intentions of their predictions, they would be violating the basic ideal of honesty that is constitutive of science in democratic societies (Douglas, 2009). On the other hand, if they are transparent about the nature of their predictions, policymakers will understandably be too mistrustful to use them to project the likely consequences of the various policy options. At the street level, caseworkers of different moral dispositions similarly will not be able to appeal to the predictions in disputes among themselves, since these are expressions of a particular moral disposition. This problem also attends the next set of approaches to the problem of performativity: by baking in a particular normative stance, pragmatic encroachment renders predictions helpless to inform disputes about the goals of allocation.

5.3 Decision-focused and policy learning

Performativity in prediction-allocation problems is an expression of a tension between the epistemic goals of accurate prediction and the normative goals of allocation. The methods we discuss in this section resolve the tension by directly recommending *decisions* expected to be the most conducive to a pre-specified normative goal. Decision-focused learning (Wilder et al., 2019; Mandi et al., 2024) and policy learning (Manski, 2004; Athey and Wager, 2021) directly output the allocation decisions which are expected to optimize a favored objective. Although predictions are made 'along the way,' they receive an instrumental interpretation and are not meant to be accurate from an epistemic perspective.

Unlike endogenization strategies, both decision-focused and policy learning issue decisions that support normative allocation principles, so long as you endorse the respective underlying allocation principle. Unlike optimal steering, both approaches are forthright about their recommendations. In their emphasis on trading certain kinds of errors, these approaches have philosophical sympathies with inductive risk arguments: large prediction errors that nevertheless recommend the best decision are preferable to small errors that do not. Unlike inductive risk arguments, they reverse the lexicographic preference for predictive accuracy in favor of non-epistemic consequences. Thus, decision-focused and policy learning are technically sophisticated expressions of the conviction that, in prediction-allocation problems, the pragmatic rightly (and dramatically) encroaches on the epistemic. The shift in focus comes at a cost. Policymakers and street-level bureaucrats who do not endorse the pre-specified goal will not know how to interpret disagreements about the correct course of action. Moreover, they will not be able to use the algorithmic recommendations to inform their own decision-making or resolve normative disputes with their colleagues. For the sake of brevity, we address the following discussion to decision-focused learning, but, except for the implementation of the learning algorithm, our points apply *mutatis mutandis* to policy learning as well.

In the standard two-stage approach to prediction-allocation problems, one first attempts to make accurate predictions. In any real-world application, we cannot expect perfectly accurate predictions. Moreover, prediction errors at the first stage will not translate straightforwardly into allocation errors at the second. Some prediction errors are epistemically large but pragmatically benign. Others are epistemically small but pragmatically vicious. Suppose that, in Pneumonia, we made admissions decisions simply by thresholding predictions of mortality risk. Then, overestimating the risk of asthmatics is pragmatically benign, since one would nevertheless make the correct admission decision. On the other hand, even small underestimates of risk can be vicious if they push someone who ought to be admitted below the admission threshold. This asymmetry in the relevance of errors is not respected by the usual "epistemic" loss function, which encourages predictions that are, on average, close to the true risk.

Decision-focused learning is motivated precisely by cases in which minimal predictive error does not ensure optimal allocation (Wilder et al., 2019; Mandi et al., 2024). Thus, decision-focused learning departs from a similar starting point as the inductive risk tradition: not all errors are equal and, even in our most epistemic activities, we should prefer to make the errors with the best non-epistemic consequences. Breaking from the two-stage procedure, decision-focused learning recommends that we train an end-to-end prediction-allocation model. Given a social welfare function, a particular allocation results in a certain level of welfare. In the idiom of deep learning, a decision-focused learning procedure iteratively propagates the welfare loss (rather than the prediction loss) backwards to update the parameters used to make predictions. After several rounds of this learning process, predictions become tailored to the chosen social welfare function. In other words: the pragmatic encroaches on the epistemic. If everything goes well, this delivers a policy for making allocations that optimizes our favorite welfare function. Crucially, however, it will not give us a sense of how good this policy is: for that we would need not only apt, but *accurate* predictions. But our predictions are now highly specialized, not to accurately predict outcomes, but to identify good candidates for intervention.

When applied to the thresholding policy in Pneumonia, decision-focused learning would likely result in a bimodal distribution of risk predictions, separating "predictions" of mortality risk from the decision threshold (Verma et al., 2023). Although inaccurate from an epistemic perspective, such exaggerations give the kind of clear signals that facilitate optimal allocation. However, for any particular patient we would not be able to tell if they have been allocated because they are just above, or well above, the threshold. This is because, unlike the inductive risk tradition, decision-focused learning lexicographically favors the non-epistemic.

At the policy level, decision-focused and policy learning fail to support rational deliberation among policy makers that hold different normative commitments. Recall the setting of Unemployment. Suppose that, among themselves, hawks and doves manage to agree on precise numerical operationalizations of their utilitarian and prioritarian commitments. Budget constraints entail that they can train only one hundred workers. Adopting a policy-learning or decision-focused learning approach, they arrive at allocation policies which are optimal from their respective normative points of view, f_u and f_p. If they are very lucky, f_u and f_p are identical, and they can suspend further debates. But typically, they will differ and some political compromise will be required. Next to normative debates about the appropriate allocation principle, finding such a compromise requires knowledge about the consequences of competing proposals. One needs to know whether one can live with the consequences of a proposed compromise. Now their troubles have begun, since neither can utilitarians tell how bad f_p is from their perspective, nor can prioritarians tell how bad f_u is from theirs. In a batch of recently unemployed workers, f_u is optimized for identifying best responders: the hundred workers for whom training would make the largest difference in employment outcomes. The prioritarian f_p is optimized for identifying the neediest cases: the hundred workers who would have the worst outcomes if they received no training. But since their predictions are specialized to optimize their preferred allocation principle, rather than predicting the consequences of allocation policies, they are not able to reliably predict the distributional consequences of their counterpart's (or their own) proposal. For the utilitarian, it is unclear whether the best responder is a significantly better responder than the worker in the hundredth (or five hundredth) position. Similarly, for the prioritarian, it is unclear whether the neediest case is significantly needier than the worker in the hundredth (or five hundredth) position. Neither can compare the consequences of training the hundred neediest cases to the consequences of training the hundred best responders. Discourse is stymied because, by baking substantive normative goals into the predictions, they have deprived themselves of the resources necessary to adjudicate their dispute.

Debates among policymakers rarely result in precise instructions for street-level decisions. Street-level bureaucrats, whether they work at a public employment service or at a hospital intake, must mediate between high-level policies on the one hand, and individuals on the other. This requires them to balance the objectives of "fairly implementing policy across all members of the political community, enforcing the state's directives, and helping those in need" (Vredenburgh, 2023). Balancing these objectives requires the exercise of bureaucratic discretion. The precise balance will depend on the caseworker's moral disposition, which she develops in the course of her work with clients and in discussions with her colleagues. Zacka (2017) provides a taxonomy of bureaucratic dispositions: 'caregivers' are eager to help those in need; 'enforcers' to sanction rule-breaking; and 'indifferent' types to process cases quickly and efficiently. Each disposition thrives in some cases and struggles with others. Especially hard cases require a reasoned discourse between competing dispositions. To avoid organizational pathologies, Vredenburgh (2023) argues that it is best to maintain a mixture of moral dispositions, so that each can sharpen itself in discourse with the other. Replacing bureaucrats with machines, Vredenburgh argues, would impose a moral monoculture, with all the weaknesses of the favored disposition.

We must expect that hard cases exist for any statistical method used to inform allocation decisions. For example, the results of the Fragile Families Challenge indicate that, despite having access to a large and high quality survey, none of 160 different modeling approaches achieved high accuracy in predicting outcomes like GPA (Salganik et al., 2019). Importantly, all models made errors on the same subset of families (Lundberg et al., 2024). This suggests that there are at least some cases that require sensitive balancing of different normative dispositions. Perhaps any particular way of balancing moral dispositions is possible to implement algorithmically. But no perfect way of balancing has yet been identified and, if we do not allow competing dispositions to continue to develop, no better way will emerge. Taken together, this suggests that while automated decision-making might result in good outcomes in typical cases, there are countervailing considerations in favor of bureaucratic discretion.

By directly outputting allocations, decision-focused and policy learning position themselves in opposition to bureaucratic discretion. Moreover, they undermine discourse among bureaucrats and frustrate the development of moral dispositions. A doveish caseworker, who places greater emphasis on helping those in need, will not be able to sharpen her judgment if all she receives is a list of the hundred neediest cases. A hawkish caseworker, who places greater emphasis on the efficient use of public funds, will not be able to sharpen her judgment if all she receives is a list of the hundred best responders. Moreover, bureaucrats will not be able to appeal to algorithmic outputs in disputes among themselves. Suppose Harry the hawk and Doris the dove disagree on whether to send Wallace the worker to training. Harry claims it would be a waste of resources, since Wallace will not find employment no matter what they do. Doris disagrees, claiming that with a bit of help, Wallace could find good work. Harry appeals to f_u: Wallace does not appear in the top hundred best responders. Doris appeals to f_p: Wallace is among the neediest cases. But just because he is a neediest case, does not mean training would make any difference to his prospects. And although Wallace is not among the best responders, that does not mean he would not respond well. Perhaps he is no worse than the worst best responder or, if the top one-hundred-and-one are equally good responders, even the best. These strategies simply do not provide the factual basis for street-level bureaucrats to discuss, critique and learn from cases. In the following, we argue that accurate estimates of counterfactual outcomes do provide the necessary support.

6 Potential Outcomes and the Pragmatics of Prediction

According to the standard diagnosis, performativity accounts for what has gone wrong in Unemployment and Pneumonia. The standard prescriptions require *managing* performativity. In this section, we give an alternative diagnosis: standard statistical predictions give rise to perverse effects because they are *infelicitous*, or pragmatically ill-formed, in prediction-allocation contexts. Rather than managing the wayward effects of infelicitous predictions, we should rather make counterfactual predictions that provide unambiguous guidance for allocation decisions.

We have seen in section 5.2 that predictive expressions like 'This patient is high risk.' are ambiguous, admitting of several interpretations, each with a different bearing on action. We now set out these interpretations more systematically. As before, we assume that the prediction is meant to inform a decision of some urgency and, for the sake of exposition, that the decision is binary: either some intervention is made (the worker receives training) or not (the patient is sent home). In such situations, predictions typically have a contrastive character, comparing the relative benefits of action and inaction. In Pneumonia, possible readings of the predictions are:

Molliative. This patient is low risk, whatever you do.
Preemptive. This patient is low risk, unless you intervene.
Fatalistic. This patient is high risk, whatever you do.
Hortative. This patient is high risk, unless you intervene.

In prediction-allocation problems, it is difficult to make predictions that are not interpreted comparatively. For example, if we say 'This patient is high risk, if you don't admit her.' there is a strong implicature that her odds are better if she gets intensive treatment. One would have to explicitly cancel the hortative interpretation by elaborating: 'But she is no better off even if you do admit her.' Why is it hard to avoid the comparative reading? Because, in these situations, predictions are directly action-guiding and therefore in the business of comparing the consequences of available acts. In contrast, standard statistical predictions are properly interpreted with a contemplative force. Given a description of an individual, they prophesy her likely outcome, in part by *guessing* what kind of treatment she will receive:

Contemplative This patient is high (low) risk, if you proceed as usual.

'Proceeding as usual' here means giving the patient the treatment that is usually given in cases like hers. Contemplative predictions are the natural idiom of gamblers and insurers, but they are infelicitous in the kinds of situations we are considering. At the hospital intake, it is a mistake of pragmatics to say 'This patient is high risk, if you proceed as usual.' Such an utterance sins against the Gricean (1975) maxims of quantity (be as formative as necessary) and manner (avoid ambiguity). It is natural to respond 'What do you mean by *as usual*?' or 'Why are you guessing what we might do, rather than helping us decide what we should do?' Because contemplative predictions are infelicitous in allocation situations, it is natural for the hearer to give them a different interpretation. In Pneumonia, a contemplative prediction of low risk is mistaken for a molliative one, giving rise to a self-defeating prophecy.[11] In Unemployment, a contemplative prediction of high risk is mistaken for a fatalistic one, giving rise to a self-fulfilling prophecy.

Pragmatically well-formed predictive expressions must account for the contrastive character of decision-relevant predictions. What are needed are accurate predictions of the counterfactual outcomes of the available courses of action e.g., if we intervene (Y^1) and if we do nothing (Y^0).[12] We adopt the notation of potential outcomes (Imbens and Rubin, 2015), a dominant formalism for causal inference.[13] Instead of predicting only the actual outcome, Y, we now target the vector (Y^0, Y^1) of counterfactual outcomes. Following Schroeder (2021) and Fuller (2021), we call such counterfactual predictions *projections*, in contrast with *forecasts* of actual outcomes. Note that we use 'prediction' in a generic way: forecasts and projections are species of the predictive genus. Standard statistical forecasts, attempting to guess what treatment the patient will receive, estimate the mixture:

$$Y = Y^1 D + Y^0(1 - D),$$

where D is a binary variable representing the action that is taken. This, ultimately, is what makes them ambiguous.[14]

In Pneumonia, asthmatics have low estimates of Y, which in their case is approximately equal to Y^1. From this, it is erroneously inferred that Y^0 is also low. If modelers report an estimate of Y^0, rather than Y, physicians can simply admit patients if the estimate of Y^0 is greater than a certain threshold: self-defeating prophecies of low risk are avoided. In Unemployment, historically disadvantaged workers have high estimates of Y, which in their case is approximately equal to Y^0. From this, it is erroneously inferred that Y^1 is also high. If modelers reported both Y^0 and Y^1, self-fulfilling prophecies of high risk would be avoided. These projections are not infelicitous because they make explicit how outcomes depend on our decisions. Moreover, it is straightforward to translate comparative predictions into statements about counterfactuals and vice-versa. Hortative predictions assert that Y^0 is appreciably larger than Y^1; fatalistic ones that Y^0 and Y^1 are both high.

We assume that counterfactual predictions, i.e. projections, do not themselves induce appreciable performative effects (Ortmann, 2025). Why should this be the case? Of course, the act of reporting a projection $(\hat{Y}^0, \hat{Y}^1)$ is *intended* to have a causal effect on the decision D and the outcome Y. What we must rule out is that the act of reporting the projection $(\hat{Y}^0, \hat{Y}^1)$ performs the true counterfactuals (Y^0, Y^1). Arguably, prediction-allocation problems can be expected to be robust to these kinds of second-order performative effects because, in these settings, predictions only influence *which* decision is made. We can think

[11] Tal (2023) makes a similar diagnosis of Pneumonia, but his suggestion misfires: admissions decisions require estimates of outcomes under outpatient, rather than ideal, treatment.

[12] Mishler (2019) similarly recommends the use of potential outcomes in recidivism prediction. Khosrowi and van Basshuysen (2024) argue that recidivism predictions must meet the requirement of "Explainability-in-Context". Our proposal also meets this requirement.

[13] Similar points can be made in the idiom of causal DAGs (Spirtes et al., 2000; Pearl, 2009).

[14] The framework can be extended to multiple treatments and our argument applies likewise.

of only two ways in which projections can perform counterfactuals: either the projection changes the *version* of the treatment ultimately administered, or it changes the *effect* of the treatment. In the setting of Pneumonia, the former kind of effect may occur if, when physicians are told that the patient will do well under inpatient treatment, they become complacent and administer worse treatment after admission. The latter kind of effect may occur if the patient overhears the prediction and becomes so alarmed that he becomes less responsive to treatment. These effects are ruled out under the (usually realistic) assumptions that (a) decision-makers are not the ones administering treatment (teaching classes, providing care) and (b) projections are either private to decision-makers or communicated with appropriate sensitivity. The first assumption ensures that reporting projections to decision-makers does not change the version of treatment and the second that it does not change the effect of treatment. Thus, although we cannot rule out higher-order performative effects *a priori*, we assume that at least here they are not appreciable: although projections $(\hat{Y}^0, \hat{Y}^1)$ perform actual outcomes Y, they do not perform counterfactual outcomes (Y^0, Y^1). Therefore, modelers should sincerely report their best estimates of counterfactual outcomes, without fearing unintended performative effects.

We argue that, once we go counterfactual, projections should be evaluated largely on epistemic grounds of accuracy. In this way, projections can be assimilated to value-neutral epistemic tools, whose purpose is to provide decision-makers with relevant information. Vredenburgh (2023) denies this possibility: since algorithms are always value-laden, they can never be evaluated on purely epistemic standards; an algorithm must always remain a “site of justice.” However, it is significantly less dangerous to apply purely epistemic standards to the evaluation of projections than forecasts. Forecasts extrapolate into the future the allocative policies of the past: they are therefore laden with the understandings and values which animated historical policy. Statistical forecasts are *reifying*: past institutional practices are taken for ‘natural laws’ to which we might accommodate ourselves, but which we cannot alter (Lukács, 1971, 87). Forecasts obscure the fact that these practices, and the social statistical regularities which arise from them, are at least partly under our control. Projections are not so badly reified as forecasts. Necessarily, they estimate counterfactual outcomes only under a small set of possible interventions and, typically, only those that have been attempted historically. Statistical methods must, by their nature, encounter difficulties in projecting the consequences of interventions that have never been tried. In this way, projections also extrapolate into the future the allocative practices of the past. However, they make explicit that outcomes are, at least in part, an expression of our own decisions and that, were we to decide differently, outcomes would change accordingly.

We may want to predict counterfactual outcomes, but is this actually feasible? Grote and Buchholz (2024) argue that, in face of the difficulties of projection, we should rather adopt standard forecasting methodologies. They point to a class of “prediction policy problems” (Kleinberg et al., 2015) for which forecasts are well-suited because predictive acts are independent of the predicted states.[15] Although projections pose methodological difficulties that do not arise in forecasting, a progressive methodological research program has arisen to address these challenges (Fischer-Abaigar et al., 2024). We do not deny the existence of prediction policy problems. With enough cleverness (which will typically

[15] Weather forecasts help decide whether to bring an umbrella, but do not change the weather.

also require some amount of explicit counterfactual reasoning) it is sometimes possible to effectively inform allocation decisions with forecasts. On the other hand, Sharma and Wilder (2024) argue, based on empirical case studies, that it is usually more effective to make targeting decision based on projections directly, rather than try to make do with forecasting actual proxies. As we have illustrated with our examples, perverse performative effects are the price of misdiagnosing a prediction-allocation as a prediction-policy problem. How often this price will have to be paid is an empirical question, that we do not attempt to answer here.

While statistical forecasts have long been individualized, methodologies for counterfactual inference have, until recently, focused on estimating *average* treatment effects i.e., the effect of treating everyone. In the setting of Unemployment, knowing that job training reduces the risk of long-term unemployment on average, though reassuring, does not necessarily help with individual allocation decisions. Recent methodological advancements such as *doubly-robust machine learning* allow for the estimation of heterogeneous treatment effects and potential outcomes at the group and *individualized* level (Bang and Robins, 2005; Belloni et al., 2017; Chernozhukov et al., 2018; Knaus, 2022).[16] These methods make potential outcomes not only conceptually adequate but also a methodologically feasible alternative target of prediction.[17]

In this section, we have offered an alternative diagnosis of the problem of performativity: difficulties arise due to the infelicity of forecasts, rather than the fact that they have a causal effect on outcomes. Going counterfactual clears up these ambiguities. However, lack of ambiguity only establishes an advantage for projections over standard statistical forecasts, not over any of the alternatives discussed in Sections 4 or 5. Ultimately, our argument for accurate projection is a normative one: accurate projection, unlike pragmatic encroachment strategies, answers to the demands of communicative rationality.

7 Discourse-Supportive Predictions

Algorithmic allocation policies empower some and impose demands on others. On a broadly liberal-democratic conception, these policies can enjoy legitimacy only if they can be justified by appeal to principles and arguments that command broad and reasoned acceptance. In light of the previous discussion, we propose two criteria of adequacy for predictions that are intended to inform allocation decisions: they should be both *decision*- and *discourse*-supportive.

The two criteria express the respective requirements of *practical* and *communicative* rationality. Following Weber, Habermas trisects practical rationality into (1) the choice of effective means for given ends; (2) the choice of appropriate ends in light of a system of values, and (3) the degree to which choices of means and ends are oriented to a system of values (1984, 172). A greater degree of practical rationality secures a greater degree of individual autonomy i.e., a greater degree of "independence from limitations imposed by the

[16] In medical trials, related methods like covariate adjustment are used for this problem (Morris et al., 2022).

[17] Like any method, doubly-robust machine learning relies on assumptions, like unconfoundedness, whose plausibility must be justified case-by-case. The evaluation of individualized estimates is also an active field of research (Knaus et al., 2020; Curth et al., 2024).

contingent environment on the self-assertion of subjects acting in a goal-directed manner" (Habermas, 1984, 15).

In Unemployment, a policy dove is practically rational to the extent that she aptly expresses her prioritarian convictions in a welfare function and adopts effective algorithmic means to find the allocation that is optimal from her perspective. Since this is precisely what policy- and decision-focused learning aim at, these techniques answer to the demands of practical rationality. But we have also seen that neither method increases the scope for the consensual resolution of normative conflicts, since they do not allow policymakers with different normative convictions to adjudicate their disputes. By taking for granted a favored normative allocation principle, the predictions fall short of supporting political deliberation. Importantly, the necessary political decisions are not given at the outset of modeling, but are themselves the outcomes of negotiation processes which require information about the likely consequences of various choices. To support such processes we need predictions which are not merely practically, but *communicatively* rational.

In contexts of communication, only those persons count as responsible who "can orient their actions to intersubjectively recognized validity claims" (Habermas, 1984, 14). Communicative rationality is not only a matter of the effective choice of means for given ends, but rather of acting on the basis of norms that can be justified to others. Thus, a greater degree of communicative rationality also expands autonomy, in the sense of the "scope for unconstrained coordination of actions and consensual resolution of conflicts" (Habermas, 1984, 15). By analogy to the distinction between practical and communicative rationality, we distinguish between *decision*- and *discourse*-supportive predictions. We say that

> a prediction is Decision Supportive for an agent if it allows her to accurately evaluate the options she faces in light of her own goals and values.

Decision-supportive predictions secure a greater degree of autonomy by creating greater scope for goal-directed self-assertion, whether for a number of individuals interacting strategically, or for a single individual acting alone. But in communicative settings, several agents must discuss which of a number of options they collectively face 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 option is best from all perspectives but, if this is not the case, it is important to be able to evaluate whether the best policy from one perspective is not also acceptable from the perspective of another, or whether some compromise can be found which is acceptable from all perspectives. We say that

> a prediction is Discourse Supportive for the agents involved in a discourse if it allows each to accurately evaluate all the options they collectively face in light of her own goals and values, as well as the goals and values of the other participants.

In the preceding, we have rejected the various proposals for *managing* performative predictions, partly because they threaten the legitimacy of the policy-making process. Endogenization and pragmatic encroachment either illegitimately arrogate to modelers the prerogative to determine the normative goals of allocation, or, by framing modelers as the passive consumers of values imported from elsewhere, abdicate the responsibility to contribute to the process by which normative goals are set. At the street level, these proposals set themselves up against bureaucratic discretion, tending to lower the level of discourse which bureaucrats hold among themselves and subvert the development of their moral dispositions. Inspired by theorists of public reason, we argue that modelers ought to deliver the kind of predictions that are conducive to arriving at (the strongest feasible) reasoned consensus. In a slogan: modelers should deliver predictions that can serve as public reasons.

At the minimum, policy deliberation should be oriented to achieving reasoned consensus among the parties to the debate. *Strong consensus* would require all parties to endorse a policy for the same reasons (Habermas, 1996, 339). In view of the likely disagreements about the normative goals of allocation, strong consensus is an ambitious goal for policy deliberations. *Weak consensus* does not require that each party shares the same reason for endorsing a policy, only that each person's justifications for endorsing it should depend on *public* reasons, i.e. reasons that all parties to the deliberation can accept as valid considerations that provide a plausible basis for accepting the proposal (Quong, 2022). The kind of consensus reached in a fractious policy debate is likely to be a weak consensus. Parties to the debate may not agree on the normative grounds on which they endorse a scheme of allocation—reasonable people may disagree e.g., about the appropriate degree of emphasis to put on helping the worst-off, as opposed to raising as many as possible above a minimum threshold, or raising the average level. Nevertheless, the *predictions* on which policymakers base their decisions should answer to high standards of publicity. At least these should be something all parties to the deliberations can accept as a plausible basis for accepting the proposal.

In the case of Pneumonia, neither standard statistical predictions nor endogenized predictions are decision-supportive, since they undermine the intersubjectively valid goals of hospital admission. Methods like decision- or policy-focused learning will typically do better on this score. In the context of Unemployment, the decision-focused predictions of f_p are indeed decision-supportive for Doris the dove, since she endorses the pre-specified goals for which the predictions are optimized. The prioritarian predictions of f_p allow Doris to identify Wallace as a neediest case and enroll him into the training program that she believes will improve his prospects. By identifying the most effective means towards her prioritarian ends, f_p removes epistemic difficulties that would otherwise constrain Doris in the assertion of her moral disposition. However, the predictions of f_p are not discourse-supportive, since they would not allow Doris to adopt the perspective of Harry the hawk, even if she wanted to. On the other hand, predictions that accurately predict Wallace's counterfactual outcomes with and without training would allow Doris to make such inferences as 'Even Harry the hawk should endorse enrolling Wallace in training, since it would significantly improve his prospects.' or 'This decision will be difficult to justify to more hawkish colleagues, since training probably won't make a large difference to Wallace's prospects.' Such projections allow Doris to distinguish situations in which value conflicts lead to unavoidable disagreements about the best course of action and those in which caseworkers with different moral dispositions can nevertheless reach a (weak) consensus on the best course of action. This allows Doris to justify herself and sharpen her moral disposition in discourse with her colleagues. Crucially, it promotes bureaucratic decision-making oriented to intersubjective validity.

At the policy level, accurate counterfactual predictions promote deliberation oriented to identifying policies which can serve as the basis of a weak consensus. Recall that, were policy hawks and doves to rely on the predictions of f_u and f_p respectively, they would not be able to discern whether the policy proposals of their counterparts are acceptable from their own perspectives. In such a situation, policymaking is prematurely reduced to a battle of wills. In their modeling exercise, Zezulka and Genin (2024) estimate the counterfactual outcomes with and without training for each individual in a batch of recently unemployed workers. Then, they project the consequences of a utilitarian and prioritarian allocation scheme on the overall level of unemployment and the gender reemployment gap. They find that a doveish allocation, which assigns the highest-risk workers to training, narrows the gender gap without increasing the level of overall unemployment, because training is more effective for higher-risk workers.[18] Their projections identify a policy that could achieve weak consensus among prioritarians interested in helping the neediest and utilitarians interested in minimizing the average rate of long-term unemployment. Such an analysis would not be possible with a pragmatic encroachment strategy. Of course, not all policy disputes can be resolved in a way that is satisfactory to all involved. But in the absence of accurate counterfactual predictions, policymakers would not be able to project the likely consequences of proposed policy compromises—the search for a weak consensus would be thrown back on intuitive political hunches.

Is it feasible to provide counterfactual predictions—e.g., estimates of potential outcomes—that are discourse supportive? Indeed, such predictions are not without precedent, even if they put high demands on estimation strategies. Suppose that, prior to modeling, discourse participants agree on a "policy space" of technically feasible and normatively reasonable allocation rules and each participant operationalizes their normative commitments in a precise welfare function. A *performative omnipredictor* (Kim and Perdomo, 2023) allows one to (1) identify the optimal member of the policy space for each of the pre-specified welfare functions and (2) for any pair consisting of a welfare function and a feasible policy, accurately project the expected outcomes and welfare level induced by the policy. The first and second conditions respectively ensure that a performative omnipredictor is decision- and discourse-supportive. Kim and Perdomo (2023) show that performative omnipredictors can be efficiently constructed from randomized trial data. It is likely that these conditions can be weakened, allowing performative omnipredictors to be constructed from the kind of (non-experimental) data that is typically available in policy contexts. While Kim and Perdomo (2023) do not make this connection explicitly, their performative omnipredictors are estimates of potential outcomes: for each decision that can be made about an individual, they estimate the induced individual distribution of outcomes.[19] In other words, performative omnipredictors do as we suggest: they provide estimates of individual, counterfactual outcomes.

[18] This cannot always be expected. Based on data from randomized controlled trials, Sharma and Wilder (2024) argue that treatment effects are often not monotonic in the predicted risk.

[19] Doubly-robust machine learning provides *unbiased* estimates of potential outcomes, but unbiasedness alone does not ensure performative omniprediction. Multi-accurate and -calibrated methods (Kern et al., 2024; Whitehouse et al., 2024) are promising, because they yield omnipredictors in static settings (Gopalan et al., 2023; Okoroafor et al., 2025).

We have proposed two adequacy conditions for predictions that inform the allocation of social goods: they ought to be *decision-* and *discourse-* supportive. The former is a demand for practical rationality: predictions must support decision-makers in achieving their respective goals. The latter is a demand for communicative rationality: predictions must respect the discretionary space of street-level bureaucrats as well as the democratic demands of policy discourse. Our proposal is in service of a conception of statistical modeling in a democratic spirit, in which modelers treat the consumers of predictions as fellow citizens whose autonomy deserves respect.

8 Conclusion: Ideology and the Value-Free Ideal

In prediction-allocation problems, standard statistical predictions give rise to undesirable self-defeating and self-fulfilling prophecies. These pernicious effects are typically blamed on performativity. Methodological proposals are made either to endogenize performative effects or to exploit them for desirable social ends. We reject these proposals because they do not answer to the demands of communicative rationality. Instead, we offer an alternative diagnosis: standard statistical predictions have undesired effects because they are infelicitous in prediction-allocation contexts. Once we go counterfactual, this performativity no longer poses a significant difficulty. The best service social statistics can render to policy discourse and to street-level bureaucrats alike is to provide accurate, counterfactual predictions. Predictions that are both *decision-* and *discourse*-supportive can provide the kind of public reasons that might form the basis of legitimate consensus. Our contribution can be seen as a plaidoyer for a version of the value-free ideal. This ideal is not an end in itself. Rather, it is intended as a proposal for how social statistics might support, rather than distort, public discourse and democratic decision-making.

We admit that, in some of its guises, the value-free ideal is ideological, in the pejorative sense of rendering technical what is properly political (Habermas, 1970). Standard contemplative predictions do have such an ideological effect: they misrepresent the results of past political choices and institutional practices as inevitable social-statistical laws to which we might accommodate ourselves, but which we cannot change. In this work, we have attempted to articulate a version of the ideal which avoids reification. A version that makes clear, within the bounds of current methodological possibility, what we must suffer and what is in our power to change.

Acknowledgments and Disclosure of Funding

We thank Thomas Grote, Markus Ahlers, Emily Sullivan, Kate Vredenburgh, Remco Heesen, Bele Wollesen, Donal Khosrowi, Moritz Hardt, Sara da Piedade Gomes, and Rabanus Derr as well as three anonymous reviewers for helpful discussions and feedback.

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Sebastian Zezulka.

References

Eva Achterhold, Monika Mühlböck, Nadia Steiber, and Christoph Kern. Fairness in Algorithmic Profiling: The AMAS Case. *Minds and Machines*, 35(1), 2025. doi: 10.1007/s11023-024-09706-9.

Anna Alexandrova. Can the Science of Well-Being Be Objective? *The British Journal for the Philosophy of Science*, 69(2):421–445, 2018. doi: 10.1093/bjps/axw027.

Doris Allhutter, Florian Cech, Fabian Fischer, Gabriel Grill, and Astrid Mager. Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective. *Frontiers in Big Data*, 3, 2020. doi: 10.3389/fdata.2020.00005.

Susan Athey and Stefan Wager. Policy Learning With Observational Data. *Econometrica*, 89(1):133–161, 2021. doi: 10.3982/ecta15732.

John Langshaw Austin. *How to Do Things With Words*. Harvard University Press, Cambridge, Mass., 1962.

Heejung Bang and James M. Robins. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61(4):962–973, 2005. doi: 10.1111/j.1541-0420.2005.00377.x.

Philippe Van Basshuysen, Lucie White, Donal Khosrowi, and Mathias Frisch. Three Ways in Which Pandemic Models May Perform a Pandemic. *Erasmus Journal for Philosophy and Economics*, 14(1), 2021. doi: 10.23941/ejpe.v14i1.582.

Fabian Beigang. Reconciling Algorithmic Fairness Criteria. *Philosophy & Public Affairs*, 51(2):166–190, 2023. doi: 10.1111/papa.12233.

A. Belloni, V. Chernozhukov, I. Fernández-Val, and C. Hansen. Program Evaluation and Causal Inference With High-Dimensional Data. *Econometrica*, 85(1):233–298, 2017. doi: 10.3982/ecta12723.

Reuben Binns. Algorithmic Accountability and Public Reason. *Philosophy & Technology*, 31(4):543–556, 2017. ISSN 2210-5441. doi: 10.1007/s13347-017-0263-5.

Liam Kofi Bright. Du Bois' democratic defence of the value free ideal. *Synthese*, 195(5): 2227–2245, 2017. doi: 10.1007/s11229-017-1333-z.

Roger C. Buck. Reflexive Predictions. *Philosophy of Science*, 30(4):359–369, 1963. doi: 10.1086/287955.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15. ACM, 2015. doi: 10.1145/2783258.2788613.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. doi: 10.1111/ectj.12097.

Gabriele Contessa. On the mitigation of inductive risk. *European Journal for Philosophy of Science*, 11(3), 2021. doi: 10.1007/s13194-021-00381-6.

Gregory F. Cooper, Constantin F. Aliferis, Richard Ambrosino, John Aronis, Bruce G. Buchanan, Richard Caruana, Michael J. Fine, Clark Glymour, Geoffrey Gordon, Barbara H. Hanusa, Janine E. Janosky, Christopher Meek, Tom Mitchell, Thomas Richardson, and Peter Spirtes. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine*, 9(2):107–138, 1997. doi: 10.1016/s0933-3657(96)00367-3.

Amanda Coston, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 582–593, New York, NY, USA, 2020. Association for Computing Machinery. doi: 10.1145/3351095.3372851.

Alicia Curth, Richard W. Peck, Eoin McKinney, James Weatherall, and Mihaela van der Schaar. Using Machine Learning to Individualize Treatment Effect Estimation: Challenges and Opportunities. *Clinical Pharmacology & Therapeutics*, 115(4):710–719, 2024. doi: 10.1002/cpt.3159.

Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD04, pages 99–108. ACM, 2004. doi: 10.1145/1014052.1014066.

Heather Douglas. Inductive Risk and Values in Science. *Philosophy of Science*, 67(4): 559–579, 2000. doi: 10.1086/392855.

Heather Douglas. *Science, Policy, and the Value-Free Ideal.* University of Pittsburgh Press, Pittsburgh PA, 2009.

W.E.B. Du Bois. The Study of the Negro Problems. *The ANNALS of the American Academy of Political and Social Science*, 11(1):1–23, 1898. doi: 10.1177/000271629801100101.

Kevin C. Elliott and Ted Richards, editors. *Exploring Inductive Risk: Case Studies of Values in Science.* Oxford University Press, 2017. ISBN 9780190467715. doi: 10.1093/acprof:oso/9780190467715.001.0001.

Jon Elster. *Local Justice: How Institutions Allocate Scarce Goods and Necessary Burdens.* Russell Sage Foundation, New York, 1992. ISBN 9780871542328.

Benjamin Eva. Algorithmic Fairness and Base Rate Tracking. *Philosophy & Public Affairs*, 50(2):239–266, 2022. doi: 10.1111/papa.12211.

Unai Fischer-Abaigar, Christoph Kern, Noam Barda, and Frauke Kreuter. Bridging the gap: Towards an expanded toolkit for AI-driven decision-making in the public sector. *Government Information Quarterly*, 41(4):101976, 2024. doi: 10.1016/j.giq.2024.101976.

Jonathan Fuller. What are the COVID-19 models modeling (philosophically speaking)? *History and Philosophy of the Life Sciences*, 43(2):1–5, 2021. doi: 10.1007/s40656-021-00407-5.

Parikshit Gopalan, Lunjia Hu, Michael P. Kim, Omer Reingold, and Udi Wieder. Loss Minimization through the Lens of Outcome Indistinguishability. In Yael Tauman Kalai, editor, *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 1–20. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. doi: 10.4230/LIPICS.ITCS.2023.60.

Herbert P. Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Volume 3: Speech acts*, pages 41–58. Academic Press, 1975.

Thomas Grote and Oliver Buchholz. Machine Learning in Public Health and the Prediction-Intervention Gap. *PhilSci Archive: Preprint*, 2024. URL https://philsci-archive.pitt.edu/23205/.

Thomas Grote, Konstantin Genin, and Emily Sullivan. Reliability in Machine Learning. *Philosophy Compass*, 19(5), 2024. doi: 10.1111/phc3.12974.

Emile Grunberg and Franco Modigliani. The Predictability of Social Events. *Journal of Political Economy*, 62(6):465–478, 1954. doi: 10.1086/257604.

Francesco Guala. *Performativity Rationalized*, pages 29–52. Palgrave Macmillan, New York, 2016. doi: 10.1057/978-1-137-48876-3_2.

Jürgen Habermas. Technology and Science as “Ideology”. In Jeremy J. Shapiro (trans.), editor, *Toward a Rational Society: Student Protest, Science, and Politics*, chapter 6, pages 81–122. Beacon Press, Boston, 1970.

Jürgen Habermas. *The Theory of Communicative Action. Volume 1: Reason and the Rationalization of Society*, volume 1. Beacon Press, Boston, 1984.

Jürgen Habermas. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Studies in Contemporary German Social Thought. MIT Press, Cambridge, MA, 1996. ISBN 0262581620.

Ian Hacking. *The Taming of Chance*, volume 17 of *Ideas in Context*. Cambridge University Press, 1990. ISBN 0521380146.

Ian Hacking. The looping effects of human kinds. In & A. J. Premack D. Sperber, D. Premack, editor, *Causal cognition: A multdisiplinary debate*, page 351–394. Clarendon Press/Oxford University Press, 1995.

Brian Hedden. On statistical criteria of algorithmic fairness. *Philosophy & Public Affairs*, 49(2):209–231, 2021. doi: 10.1111/papa.12189.

Carl G. Hempel. Science and Human Values. In Rpbert E. Spiller, editor, *Social Control in a Free Society*, pages 39–64. University of Pennsylvania Press, 1960. doi: https://doi.org/10.9783/9781512807424.

Ralph Hertwig and Till Grüne-Yanoff. Nudging and Boosting: Steering or Empowering Good Decisions. *Perspectives on Psychological Science*, 12(6):973–986, 2017. doi: 10.1177/1745691617702496.

Hannah Hilligardt. Science as public service. *European Journal for Philosophy of Science*, 14(3), 2024. doi: 10.1007/s13194-024-00607-3.

Lily Hu, Nicole Immorlica, and Jennifer Wortman Vaughan. The Disparate Effects of Strategic Manipulation. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 2019. doi: 10.1145/3287560.3287597.

Guido W. Imbens and Donald B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Richard C. Jeffrey. Valuation and Acceptance of Scientific Hypotheses. *Philosophy of Science*, 23(3):237–246, 1956. doi: 10.1086/287489.

Christoph Kern, Michael Kim, and Angela Zhou. Multi-Accurate CATE is Robust to Unknown Covariate Shifts. *Transactions on Machine Learning Research*, 2024.

Donal Khosrowi. Managing Performative Models. *Philosophy of the Social Sciences*, 53(5):371–395, 2023. doi: 10.1177/00483931231172455.

Donal Khosrowi and Philippe van Basshuysen. Making a Murderer. *American Philosophical Quarterly*, 61(4):309–325, 2024. doi: 10.5406/21521123.61.4.02.

Michael P. Kim and Juan C. Perdomo. Making Decisions Under Outcome Performativity. In Yael Tauman Kalai, editor, *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251, pages 1–15. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023. doi: 10.4230/LIPICS.ITCS.2023.79.

Philip Kitcher. *Science, Truth, and Democracy*. Oxford Studies in Philosophy of Science Series. Oxford University Press, 2001. doi: 10.1093/0195145836.001.0001.

Philip Kitcher. Science in a Democratic Society. In Wenceslao J. Gonzalez, editor, *Scientific Realism and Democratic Society: The Philosophy of Philip Kitcher*, volume 101 of *Poznań Studies in the Philosophy of the Sciences and the Humanities*, pages 95–112. Rodopi, 2011. doi: 10.1163/9789401207355_003.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction Policy Problems. *American Economic Review*, 105(5):491–495, 2015. doi: 10.1257/aer.p20151023.

Michael C. Knaus. Double machine learning-based programme evaluation under unconfoundedness. *The Econometrics Journal*, 25(3):602–627, 2022. doi: 10.1093/ectj/utac015.

Michael C. Knaus, Michael Lechner, and Anthony Strittmatter. Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal*, 24(1):134–161, 2020. doi: 10.1093/ectj/utaa014.

Matthew Kopec. A More Fulfilling (and Frustrating) Take on Reflexive Predictions. *Philosophy of Science*, 78(5):1249–1259, 2011. doi: 10.1086/662266.

Matthias Kuppler, Christoph Kern, Ruben L. Bach, and Frauke Kreuter. From fair predictions to just decisions? conceptualizing algorithmic fairness and distributive justice in the context of data-driven decision-making. *Frontiers in Sociology*, 7, 2022. doi: 10.3389/fsoc.2022.883999.

Michael Lipsky. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Service.* Russell Sage Foundation, New York, 2010.

Helen E. Longino. Gender, politics, and the theoretical virtues. *Synthese*, 104(3):383–397, 1995. doi: 10.1007/bf01064506.

György Lukács. Reification and the Counsciousness of the Proletariat. In *History and Class Consciousness: Studies in Marxist Dialectics*, page 83–222. MIT Press, 1971.

Ian Lundberg, Rachel Brown-Weinstock, Susan Clampet-Lundquist, Sarah Pachman, Timothy J. Nelson, Vicki Yang, Kathryn Edin, and Matthew J. Salganik. The origins of unpredictability in life outcome prediction tasks. *Proceedings of the National Academy of Sciences*, 121(24), 2024. doi: 10.1073/pnas.2322973121.

Donald A. MacKenzie. *An engine, not a camera.* Inside technology. MIT Press, Cambridge, MA, 2006.

Jayanta Mandi, James Kotary, Senne Berden, Maxime Mulamba, Victor Bucarey, Tias Guns, and Ferdinando Fioretto. Decision-Focused Learning: Foundations, State of the Art, Benchmark and Future Opportunities. *Journal of Artificial Intelligence Research*, 80:1623–1701, 2024. doi: 10.1613/jair.1.15320.

Charles F. Manski. Statistical Treatment Rules for Heterogeneous Populations. *Econometrica*, 72(4):1221–1246, 2004. doi: 10.1111/j.1468-0262.2004.00530.x.

Celestine Mendler-Dünner, Frances Ding, and Yixin Wang. Anticipating Performativity by Predicting from Predictions. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 31171–31185. Curran Associates, Inc., 2022.

Robert K. Merton. The Self-Fulfilling Prophecy. *The Antioch Review*, 8(2):193, 1948. doi: 10.2307/4609267.

John P. Miller, Juan C. Perdomo, and Tijana Zrnic. Outside the Echo Chamber: Optimizing the Performative Risk. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7710–7720. PMLR, 2021.

Alan Mishler. Modeling Risk and Achieving Algorithmic Fairness Using Potential Outcomes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19. ACM, January 2019. doi: 10.1145/3306618.3314323.

Tim P. Morris, A. Sarah Walker, Elizabeth J. Williamson, and Ian R. White. Planning a method for covariate adjustment in individually randomised trials: a practical guide. *Trials*, 23(1), 2022. ISSN 1745-6215. doi: 10.1186/s13063-022-06097-z.

Robert Northcott. Reflexivity and fragility. *European Journal for Philosophy of Science*, 12(3), 2022. doi: 10.1007/s13194-022-00474-w.

Princewill Okoroafor, Robert Kleinberg, and Michael P. Kim. Near-Optimal Algorithms for Omniprediction, 2025.

Jakob Ortmann. Performative paternalism. *European Journal for Philosophy of Science*, 15(2), 2025. doi: 10.1007/s13194-025-00651-7.

Judea Pearl. *Causality.* Cambridge University Press, Cambridge, second edition, 2009.

Juan Perdomo. *Performative Prediction: Theory and Practice.* Phd thesis, University of California, Berkeley, Berkeley, CA, 2023.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative Prediction. In Hal Daumé III. and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609. PMLR, 2020.

Juan C. Perdomo. Revisiting the Predictability of Performative, Social Events, 2025.

Juan C. Perdomo, Tolani Britton, Moritz Hardt, and Rediet Abebe. Difficult Lessons on Social Prediction from Wisconsin Public Schools, 2023.

Hilary Putnam. *The Collapse of the Fact/Value Dichotomy and Other Essays.* Harvard University Press, Cambridge, MA, 2002.

Jonathan Quong. Public Reason. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, Summer 2022 edition, 2022.

John Rawls. The Idea of Public Reason Revisited. *The University of Chicago Law Review*, 64(3):765–807, 1997. doi: 10.2307/1600311.

George D. Romanos. Reflexive Predictions. *Philosophy of Science*, 40(1):97–109, 1973. doi: 10.1086/288499.

Richard Rudner. The Scientist Qua Scientist Makes Value Judgments. *Philosophy of Science*, 20(1):1–6, 1953. doi: 10.1086/287231.

Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, and Sara McLanahan. Introduction to the Special Collection on the Fragile Families Challenge. *Socius: Sociological Research for a Dynamic World*, 5:237802311987158, 2019. doi: 10.1177/2378023119871580.

S. Andrew Schroeder. How to Interpret Covid-19 Predictions: Reassessing the IHME's Model. *Philosophy of Medicine*, 2(1), 2021. doi: 10.5195/pom.2021.43.

Vibhhu Sharma and Bryan Wilder. Comparing Targeting Strategies for Maximizing Social Welfare with Limited Resources, 2024.

Ali Shirali, Rediet Abebe, and Moritz Hardt. Allocation Requires Prediction Only if Inequality Is Low. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 45114–45153. PMLR, 2024.

Herbert A. Simon. Bandwagon and Underdog Effects and the Possibility of Election Predictions. *Public Opinion Quarterly*, 18(3):245, 1954. doi: 10.1086/266513.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search.* MIT Press, Cambridge, Mass, second edition, 2000.

Jason Stanley. *Knowledge and Practical Interests.* Oxford University Press, New York, 2005.

Eran Tal. Target specification bias, counterfactual prediction, and algorithmic fairness in healthcare. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23. ACM, 2023. doi: 10.1145/3600211.3604678.

Richard H. Thaler and Cass R. Sunstein. *Nudge: Improving Decisions about Health, Wealth, and Happiness.* Yale University Press, London, 2008.

Johanna Thoma. Social Science, Policy and Democracy. *Philosophy & Public Affairs*, 2023. doi: 10.1111/papa.12250.

Shresth Verma, Aditya Mate, Kai Wang, Neha Madhiwalla, Aparna Hegde, Aparna Taneja, and Milind Tambe. Restless Multi-Armed Bandits for Maternal and Child Health: Results from Decision-Focused Learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, page 1312–1320, Richland, SC, 2023. International Foundation for Autonomous Agents and Multiagent Systems.

Ulrike von Luxburg and Bernhard Schölkopf. Statistical Learning Theory: Models, Concepts, and Results. In Dov M. Gabbay, Stephan Hartmann, and John Woods, editors, *Inductive Logic*, volume 10 of *Handbook of the History of Logic*, pages 651–706. North-Holland, 2011. doi: https://doi.org/10.1016/B978-0-444-52936-7.50016-1.

Kate Vredenburgh. AI and bureaucratic discretion. *Inquiry*, pages 1–30, 2023. doi: 10.1080/0020174x.2023.2261468.

Justin Whitehouse, Christopher Jung, Vasilis Syrgkanis, Bryan Wilder, and Zhiwei Steven Wu. Orthogonal Causal Calibration, 2024.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the Data-Decisions Pipeline: Decision-Focused Learning for Combinatorial Optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):1658–1665, 2019. doi: 10.1609/aaai.v33i01.33011658.

Eric Winsberg and Stephanie Harvard. Purposes and duties in scientific modelling. *Journal of Epidemiology and Community Health*, 76(5):512–517, 2022. doi: 10.1136/jech-2021-217666.

Bernardo Zacka. *When the State Meets the Street: Public Service and Moral Agency*. Harvard University Press, Cambridge, MA, 2017.

Sebastian Zezulka and Konstantin Genin. Performativity and Prospective Fairness. *arXiv preprint arXiv:2310.08349*, 2023.

Sebastian Zezulka and Konstantin Genin. From the Fair Distribution of Predictions to the Fair Distribution of Social Goods: Evaluating the Impact of Fair Machine Learning on Long-Term Unemployment. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT'24, New York, 2024. Association for Computing Machinery. doi: https://doi.org/10.1145/3630106.3659020.