Machine Learning in Public Health and the Prediction-Intervention Gap

Thomas Grote
thomas.grote@uni-tuebingen.de

University of Tübingen
Cluster of Excellence "Machine Learning: New Perspectives for Science"

Oliver Buchholz
oliver.buchholz@hest.ethz.ch

ETH Zürich
Chair of Bioethics

Forthcoming in Durán, J. & Pozzi, G. (eds.): *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*. Cham: Synthese Library.

Abstract: This chapter examines the epistemic value of (purely) predictive ML models for public health. By discussing a novel strand of research at the intersection of ML and economics that recasts policy problems as prediction problems, we argue – against skeptics – that predictive models can indeed be a useful guide for policy interventions, provided that certain conditions hold. Using behavioral approaches to policymaking such as Nudge theory as a contrast class, we carve out a distinct feature of the ML approach to public policy problems: the ML model itself may turn into a cognitive intervention. In underscoring the epistemic value of predictive models, we also highlight the importance of taking a broader perspective on what constitutes good evidence for policymaking. Moreover, by focusing on public health, we also contribute to the understanding of the specific methodological challenges of ML-driven science outside of traditional success areas.

Keywords: Machine Learning; Public Health; Prediction; Health Economics; Algorithmic Decision-Making; Evidence-Based Policymaking.

1. Introduction

Fueled by successes in image-based diagnostics and risk-prediction models, there is ongoing enthusiasm for using machine learning (ML) models as decision-support tools in clinical medicine (Gulshan et al. 2016; Esteva et al. 2017; Hyland et al. 2020). In comparison, the adequate role of ML applications has yet to be found in the context of public health. This may seem surprising, given that health data from virtually every domain of people's lives are being collected digitally, potentially improving healthcare policies. A particular obstacle relates to the gap between

predictions and interventions: while predicting unknown outcomes from patterns of past observations can be useful, how can policymakers anticipate which interventions are most likely to change the outcome in desirable ways? Besides, ML models perform demonstrably well when they are deployed in environments that mirror their training conditions. Yet they demonstrably do not perform well beyond training conditions (Freiesleben and Grote 2023). The environment in public health contexts is typically unstable – culminating in prediction errors. In consequence, many researchers argue that purely predictive models are not the right tools to inform public health interventions (Athey 2017; Hernán et al. 2019).

To overcome the gap between predictions and interventions, the proposed solution is to either supplement predictive ML models with causal inference methods or to embrace a causal attitude towards data science problems in public health altogether (Cui and Athey 2022; Hernán et al. 2019; Schölkopf et al. 2021). While this view has become increasingly influential, it is not unequivocally shared. For example, Broadbent and Grote (2022) contend that it is misguided to impose causal constraints on ML models in epidemiology, since such constraints could prevent the discovery of novel statistical relationships that have the potential to revise the apparatus of public health concepts. However, their account sidesteps the issue of how the gap between predictions and healthcare interventions can be diminished. We do not try to settle the dispute between predictive and causal modeling cultures in public health in this chapter. Instead, we pursue the relatively modest goal of understanding how (and to what extent) purely predictive ML models can guide public health interventions.

For this purpose, we particularly draw on a study from the field of health economics by Mullainathan and Obermeyer (2022) who are using ML models as a tool to precisely characterize inefficiencies in the healthcare systems by examining how physicians' decisions deviate from

predicted risk in the diagnosis of heart attack.[1] The upshot of the study is that physicians systematically overtest low-risk patients while undertesting high-risk patients. By developing a nuanced picture of the microfoundations of physician testing behavior, Mullainathan and Obermeyer conclude that current moral hazard models, such as low-testing regimes, can have perverse effects (see also Hausman 2021). The approach taken in this study is paradigmatic for a strand of research at the intersection of ML and economics that recasts the design of optimal policies in criminal justice, social welfare, or healthcare as prediction problems (Kleinberg et al. 2015; Kleinberg et al. 2018).

That said, whereas Mullainathan and Obermeyer's (2022) modeling of physicians' decision-making allows ruling out misguided policies, it is unclear how it can translate into the design of more efficient public health interventions – other than providing physicians with algorithmic decision-support tools. This issue is closely linked to the granularity of the relevant risk-prediction model. In that respect, there are some striking discontinuities between the ML approach and other strategies of policy-design grounded in the behavioral sciences. A pertinent example here is Nudge theory. Guided by the assumption that decisions can result from two styles of information processing, the objective is to intervene in the decision architecture in a way that counteracts biases and ultimately enables people to act rationally/in desirable ways (Thaler and Sunstein 2008). By contrast, the ML approach to public health problems is likely to result in interventions that diminish the role of policymakers: the ML model sets the bar for which payoff function to maximize and for what counts as the optimal decision-rule, leaving little room for policymakers. Hence, through a careful analysis of the methodology and its underlying constraints, the hope is to gain a clearer picture regarding the scope of ML approaches to public health problems. In this vein, this chapter

[1] Although analyzing the decisions of individual physicians, the study is adopting a public-health perspective because its ultimate aim is to ground health-related policies on said individual decisions rather than on aggregate views (Mullainathan and Obermeyer 2022, p. 723).

also provides important groundwork that allows science and society alike to think more systematically about the role of ML approaches within the context of evidence-based policymaking.

The chapter proceeds as follows: Section 2 explains the gap between prediction and intervention in ML approaches to public health by pointing out various factors that might undermine the external validity of purely predictive models. Section 3 studies the methodology of the ML approach to policy problems. Of particular importance here is specifying what it means to reframe policy problems as prediction problems, as opposed to matters of causal inference. Furthermore, (dis)continuities between the ML approach inspired by behavioral science will be highlighted. Finally, Section 4 considers to what extent the ML approach to public health problems manages to overcome the gap between predictions and interventions. For this purpose, we examine the paper by Mullainathan and Obermeyer (2022) as a case-study. To carve out the opportunities and pitfalls of the ML approach, Nudge theory is taken as a contrast class. Overall, we hope that this chapter stimulates further reflection on the proper role of ML approaches in public health, while also broadening the debate on the epistemology and methodology of policymaking.

2. The Prediction-Intervention Gap

The objective in (supervised) ML is to facilitate the learning of mathematical functions that achieve high predictive accuracy on new data. Clinical medicine in particular has shown to be a beneficiary of the recent breakthroughs of ML in image-based classification and risk modeling. For instance, using measurements from multiple organ systems as input data, Hyland et al. (2020) developed an ML model that predicts circulatory failure in patients in the intensive care unit (ICU) with an accuracy of up to 90%. ICU patients cannot be monitored continuously by nurses or physicians, and it is exceedingly difficult for them to process the large quantities of data from many patients stored in electronic patient-data management systems. The risk-prediction model might therefore

act as an early-warning system for deteriorating patients, potentially enabling timely interventions and decreasing the number of false alerts.[2]

This study is exemplary for dozens of ML applications meant to be used as decision-support tools in clinical medicine. By contrast, studies on public health applications of ML are still scarce – at least when measured in the currency of publications in high-profile (medical) journals. With 'public health', we refer to the societal practice that deals with the health of a country or region, driven by goals that manifest in policy documents and regulations (Munthe 2008, p. 39; see also Verweij and Dawson 2007). For example, while many leading research groups have been eager to develop ML-based forecasting models in the course of the SARS-CoV-19 pandemic, national health institutions usually were inclined to stick to stochastic compartmental epidemiological models, such as the Susceptible-Exposed-Infectious-Recovered (SEIR) model to control for the effectiveness of social distancing policies and mask mandates.[3] There is little need to argue that powerful predictive models can be useful for evidence-based policymaking. So why is it that ML applications still play an ancillary role in the methodological repertoire of public health?

Especially when factoring out material constraints – e.g., a lack of available datasets for public health purposes – then the underlying problem can be expressed by a simple slogan: predictions are not interventions. However, behind this slogan lies a multi-layered set of issues. We follow the orthodoxy in ML and define 'prediction' as an estimation of unknown outcomes based on patterns found in past observations (Hardt and Recht 2022, p. 16). Predictions and interventions can be individuated by their distinct causal roles. Predictions are causally upstream to interventions. If a model predicts an increase in virus spread, then this might guide policies to contain the spread. By contrast, the intervention/policy sets out to change the healthcare system. For a predictive model

[2] In general, applications of this kind are at risk of hiding the ML model's ignorance with respect to one particular patient behind a high average accuracy, thereby running into a setting that resembles the well-known reference class problem. However, see Buchholz (2023b) for an analysis of this problem and of situations in which it might not be that much of a problem after all.

[3] With some exceptions, such as the city of Valencia using ML-based prediction models to forecast virus spread in real-time (Díaz-Lozano et al. 2022). For a comprehensive treatment of ML applications amidst the SARS-CoV-19 pandemic, see Syrowatka et al. (2021) for review.

to guide interventions, it must meet the conditions of robustness and control. The condition of robustness states that the model maintains its predictive performance (to a significant degree) in a deployment scenario regardless of (reasonable) changes in the probability distribution of the data (Freiesleben and Grote 2023). This is meant to counteract the threat that predictions are based on artifacts. This, in turn, facilitates trust in the model. In contrast, the condition of control states that it must be possible to extract actionable insights from the model about how and when to intervene (see also Cui and Athey 2022).

Recall the risk-prediction model by Hyland et al. (2020): One reason for its clinical utility is that it is being used in narrowly confined settings. The medical devices measuring patients' organ systems are likely to remain stable over time and ICU patient management is highly regulated. Once the predicted risk reaches a pre-defined threshold, the alarm will go off and a physician will initiate the corresponding treatment. Especially since physicians have background knowledge concerning patients' physiology and appropriate treatment options, there is a straightforward path from predicted risk to intervention. Even still, dataset shifts (e.g., the model gets applied to new demographics or policies in the clinic have changed) might induce malfunctions. However – for the most part – these are controllable and can be mitigated by retraining the model (Finlayson et al. 2021; Freiesleben and Grote 2023).

2.1 Skepticism Against (Purely) Predictive Models

In contrast to the setting studied by Hyland et al. (2020), many environments that are of concern for public health lack clear boundaries and are subject to rapid changes. The predictive performance of an epidemiological model might be impacted by seasonal changes in virus spread or changes in the population's behavior. A case in point is Google's Flu Trends (GFT), trying to predict the prevalence of flu from people's online searches. One reason that led to the cancelation of GFT was a steep increase in online searches on flu, leading the model to drastically overestimate flu prevalence. This was particularly due to flu being prominently discussed in the news in fall 2013

(Lazer et al. 2014). Moreover, some changes to the environment can be even triggered by the deployment of the model itself, e.g., when people strategically adjust their behavior to game its functionality. This issue is commonly discussed as the *performativity* of (ML-based) predictions (Perdomo et al. 2020).[4]

Further complications arise when seeking to extract actionable information from ML models to guide policies. Critics of current ML are keen to point out that ML models often do not achieve high predictive accuracy by registering meaningful variables, but by exploiting spurious associations (Pearl and Mackenzie 2019, Ch. 10; Geirhos et al. 2020). In this case, any information extracted from the model could mislead policymakers. Another problem results from the challenge to extract meaningful information from highly complex and, thus, opaque ML models in the first place – and from the question to what extent the relevant opacity can be ameliorated by making said models explainable (Creel 2020; Sullivan 2022; Boge 2022).[5] After all, policymakers do not just want to predict an outcome of interest but to understand which adjustments are necessary to steer the outcome in a desirable way. Closely related thereto, for evaluating suitable policies, it is not only important to predict a given outcome based on observed data, but also to simulate how the predicted output were to change if a new policy gets released (Athey 2017).

A widely shared view is that tackling these problems requires a different set of statistical techniques than pure predictive modeling.[6] For instance, Hernán et al. (2019) are skeptical about the premise that predictive models can inform public health policies. On their view, predictive models inform us that a certain decision must be made but they cannot guide the respective decision itself.

[4] See also Northcott (2017, p. 633) on the openness of social environments.
[5] Speaking of *relevant* opacity is meant to emphasize the insight put forward, for instance, by Buchholz (2023a) that ML models are not opaque *per se*, but different aspects of them are deemed opaque by different individuals, thus making the issue of explainability a highly granular epistemic endeavor.
[6] There are statistical techniques, for instance linear regression models, that can be used both for making predictions and – provided specific assumptions are met – for inferring causal relationships between variables. Standard ML methods like deep neural networks or support vector machines, however, can exclusively be used for the former.

Ultimately, what matters is that policymakers can tackle "what if"-questions – which, according to Hernán et al., can only be achieved by causal inference techniques (p. 49).

In response to these kinds of arguments, several authors have tried to make progress on a methodological shift that involves either amalgamating or at least complementing ML models with causal inference techniques. What is at stake here is capturing dependencies between the effect variable *X* and the outcome variable *Y* that remain invariant across different settings. As argued by Schölkopf et al. (2021), such an approach would lead to models that are more robust and thus better able to generalize to data that have been sampled under different conditions than the original training set. To make steps into this direction, Schölkopf et al. (2021) suggest abandoning the crucial – and, in fact, often unrealistic – *i.i.d.*-assumption of ML according to which all instances in the training and test data must be sampled independently from each other as well as from the identical probability distribution. Instead, they argue that one should impose weaker assumptions, for instance, that "the data on which the model will be applied comes from a possibly different distribution but involving (mostly) the same causal mechanisms" (p. 624). In practice, this could be achieved by using a more diverse training set and by employing techniques of data augmentation both of which would force the model to learn "underlying invariances or symmetries present in the augmented data distribution" (p. 626).[7]

A similar approach is pursued with the stable learning framework that has recently been proposed by Cui and Athey (2022): Instead of assuming that the training and test data come from the same distribution, the framework relies on the assumption that both might be from different so-called environments. The overall aim of stable learning then is to learn the mapping between a potentially larger number of treatment variables and the outcome, thereby determining a model that "can achieve uniformly good performance on any possible environment" (p. 112). Note that, in this context, 'performance' is taken to be *predictive* performance. This reveals how the stable learning

[7] A common example for data augmentation in the case of image classification is the transformation of training instances by, e.g., rotating images or varying a few pixels.

framework amalgamates the traditional ML paradigm with a notion of causality according to which a model that captures (parts of) the causal structure of an underlying phenomenon is robust to interventions and, thus, remains stable across different environments.

In a similar vein, Hofman et al. (2021) propose an integrative modeling framework for the purpose of reconciling predictive and explanatory modeling approaches in computational social science. Roughly, their idea is to use explanatory models to identify and estimate causal effects. The validity of the causal assumptions will then be determined by whether they enable predicting outcomes that can be considered, yet again, as being out-of-distribution relative to the training data. To map the explanatory models onto the predictive models, Hofman et al. suggest various methods, ranging from counterfactual sanity checks to sophisticated knowledge extraction techniques (see also Hinton et al. 2015).[8]

2.2. The Epistemic Value of Purely Predictive Models: A Modest Defense

The picture that emerges against the backdrop of the last section is that predictive models are ill-suited to guide public health interventions, at least unless they are supplemented by causal inference techniques. While it falls outside of this chapter's scope to further engage with the technicalities of the respective causal inference frameworks, we want to argue that it might be premature to abandon the prospect of using purely predictive models for improving public health policies.

Consider a pragmatic concern: Even though the proposals to bridge predictive models with causal inference techniques are highly persuasive, these are still theoretical frameworks. As it stands, we are not aware of any application of the stable learning framework within the context of public health/policy. By contrast, training a powerful predictive ML model can be a matter of weeks, provided that there is enough data, computational resources, and expertise. Hence, until methods that combine ML models with causal inference techniques have been fleshed out, it could be useful

[8] For a more detailed discussion of said integrative modelling framework as well as its perks and perils in the social sciences, see Buchholz and Grote (2023).

to suspend any constraints on using purely predictive models for public health purposes. A substantive argument for skepticisms against causal ML is that the relevant frameworks typically rely on (overly) strong assumptions, such as un-confoundedness in the data, which cannot be presumed for real-world settings (see also Gelman 2010).

Then, there are methodological considerations: As the environment in public health contexts is oftentimes messy, it can be challenging to identify and estimate the true causal structure. Public health is an undertheorized domain, many of the necessary assumptions to establish causal claims are difficult to test or might only be justified by large-scale randomized controlled trials (RCTs), of which there are not enough (see also Cui and Athey 2022). The lack of RCTs for public health purposes is closely tied to logistical (conducting RCTs can be very expensive and time-consuming) and ethical considerations (since they entail that one group receives worse treatment in what are often high-stakes settings). Resulting from this is the problem to delineate causal effects from confounders. The fact that explanatory models in the social sciences oftentimes lack predictive power showcases the difficulty to identify the causal structure in social environments. Indeed, it is debatable whether social environments that have a robust causal structure, holding reliably across cases, exist in the first place (Northcott 2020; 2022). And while simpler epidemiological models like the SEIR model are also not strictly speaking causal, they provide researchers and policymakers with a lot of control, since it is easy to see how manipulations of individual features affect the outcome of interest.

Moreover, although one should be cautious to not naively interpreting the function of a predictive model as indicating the discovered structure, it is reasonable to assume that if an ML model predicts social phenomena at high accuracy, then *some* structure (or at least stable parameters that can guide policy-interventions) will have been found (Mullainathan and Spiess 2017, p. 98). One particular promise of using ML models is that they enable discovering structures that so far were elusive to researchers and policymakers. Applied to the context of public health, this could include cross-connections between different kinds of data (e.g., electronic health records, mobility data, lifestyle

data, and so on) (Broadbent and Grote 2022). At least in principle, this might be valuable for developing better public health interventions.

However, a prerequisite for assessing how ML models can guide public health interventions is to test them against practical reality: Which phenomena can be predicted at high accuracy – and which cannot? How can researchers ensure the model's external validity? What are inherent methodological constraints? How can policymakers draw actionable insights from the model? What sorts of interventions might the model guide? Answering these questions requires a pronounced understanding of ML approaches to public health.

3. The Machine Learning Approach to Policy Problems

In what follows, we turn to a detailed methodological analysis of the ML approach to policy problems. The distinctive feature of this approach – receiving increased attention in economics – is that it tries to solve policy problems by using predictive algorithms, while disentangling policy choices from causal inference.

3.1. Prediction Policy Problems

According to a paper by Kleinberg et al. (2015), policy problems differ regarding their methodological requirements: If a policymaker facing a drought must decide whether to invest in a rain dance to increase the chance of rain, then causal inference techniques are necessary. What matters is whether rain dances cause rain. By contrast, if the task is to decide whether to take an umbrella to work to avoid getting wet considering the clouds at the sky, then this is not a causal inference but a prediction problem. Unlike the rain dance example, taking an umbrella has no direct effect on rain. However, to evaluate the utility of taking an umbrella to work, it is important to estimate the chance of rain. Put differently, it is unknown whether it will rain, and a corresponding prediction is needed to decide about taking an umbrella, while the causal relationship between using an umbrella and staying dry in the case of rain is known and, thus, requires no further investigation. The claim is that such prediction policy problems have been neglected so far and that ML models

can help in solving them more effectively, when compared to traditional statistical methods such as regression analysis (p. 491).

Kleinberg et al. (2015) do not formally define prediction policy problems but provide a list of illustrative examples for potential empirical applications. Among those are predicting which teacher will have the greatest value added, predicting unemployment spell length to help workers decide on savings rates and job search strategies, targeting health inspections, predicting highest risk youth for targeted interventions, or predicting the creditworthiness of borrowers in lending decisions (p. 494). If we abstract from these examples, the common structure of prediction policy problems is that there is a known loss function, and the task is to minimize error when predicting a target variable Y from input data X. To this list, we can add some conditions that need to be met whenever some problem ought to be solved by ML techniques: There must be enough data to train the ML model, the target variable needs to be narrowly confined as a precondition for keeping measurement error small, and the background conditions need to be stable over time. In consequence, prediction policy problems relate to routine tasks, as opposed to predicting black swan events. For instance, predicting the spread of a novel virus, whose characteristics are not yet well understood, clearly falls outside of their scope.[9]

To showcase the structure of prediction policy problems, consider a lending decision. Here, a bank is confronted with a binary decision whether to grant or deny a loan application based on the predicted creditworthiness of the borrower. Similarly, while the impact of hiring an additional teacher can be well estimated, hiring the *right* teacher requires predicting individual teacher quality from information available at the time of hiring (Mullainathan and Spiess 2017, p. 102). Still, things might not be as simple. A skeptic might object that assessing the suitability of a teacher for a particular school may largely hinge on a few causal factors – e.g., their professional skills and certain

[9] At a more abstract level, the underlying problem is that the remarkable success of ML models in interpolating from data is not matched in terms of extrapolation. For an intuitive explanation of the difference between interpolation and extrapolation, see also Freiesleben and Grote (2023).

psychological traits – rather than being a matter of some large-scale prediction exercise. After all, how can an ML model predict whether a teacher fits well into a particular school culture? Likewise, without knowing why certain youths are at risk, any targeted intervention might go astray. The point is that many actual examples of prediction policy problems involve more considerate choices than merely finding a decision-rule that facilitates minimizing error in a known payoff function. Somewhat pointedly formulated, it might be argued that many alleged prediction policy problems fall prey to the 'law of the instrument' (Maslow 1966): suddenly, policymakers have access to vast amounts of data and highly predictive models, with the result that everything turns into a prediction problem.

Moreover, not everything that may seem like a natural candidate for a prediction policy problem is in fact predictable. For example, to assess whether medical spending in the last year of a life is wasteful – amounting to one-quarter of a person's health-costs in total – Einav et al. (2018) used ML techniques to predict mortality among patients. The result being that less than 5% of spending is accounted for by individuals with predicted mortality above 50%. With that in mind, the unpredictability of mortality can also be deemed as evidence that medical spending at the end of the life is not wasteful. In a similar manner, a mass collaboration of 160 teams has shown low accuracy in predicting life trajectories from the Fragile Families and Child Wellbeing Study dataset (Salganik et al. 2020). However, what makes said phenomena unpredictable is not yet well-settled; either the available data proves to be insufficient, or the phenomenon of interest happens to be too complex to be operationalized as a prediction target. We will put this issue aside for now, but see Buchholz and Grote (2023) for details.

The programmatic considerations on the nature of prediction policy problems culminated in an influential study, in which Kleinberg et al. (2018) examine how ML models can be used to understand and improve judges' decisions whether to release defendants pre-trial in the US criminal justice system. Importantly, such decisions are generally assumed to solely hinge on the likelihood of the defendant to commit further crimes if they get released. Pre-trial decisions are therefore

paradigmatic of a prediction policy problem for ML. The target variable is narrowly confined, and the data are extensive, provided that there are ten million arrests in the US per year. For the study, Kleinberg et al. use a dataset of 758,027 defendants who were arrested in New York City between 2008 and 2013, including detailed information on the defendants – e.g., whether they were released pre-trial and went on committing further crimes in-case. From a methodological perspective, the basic idea of the study is to train an ML model on the dataset for the aims of identifying the optimal decision function for pre-trial decisions. The model then acts as a benchmark, measuring judges' decision-making. One result being that judges are prone to release roughly half of the defendants, predicted to be within the riskiest 1% of defendants. This even applies to strict judges, drawing additional detainees from the whole risk distribution. As an upshot, Kleinberg et al. argue that if strict judges were to decide in accordance with the algorithmically predicted risk, they could achieve the same reduction in crime by only arresting half as many defendants. (p. 240).

With that in mind, a particular merit of the study is that Kleinberg et al. (2018) are very explicit about the challenges in comparing ML-based risk models to judges' decisions: There is no data on whether incarcerated defendants would have committed crimes if they would have been released, deviance from the algorithmically predicted risk can be attributed to different motivational factors on the judge's side, and there are certain kinds of evidence available to judges that the respective ML models cannot register. To counteract these obstacles, Kleinberg et al. use different techniques from the econometrics toolkit. These will be explained in greater detail at a later point in this chapter.

3.2 Diagnosing Error in Human Judgement

While paradigmatic for the ML approach to policy problems, the study by Kleinberg et al. (2018) itself can be situated in the tradition of two research strands concerned with understanding inaccuracies in human judgement. The first one is comparative research on two distinct methods of judgment, namely clinical versus actuarial judgement (Dawes, Faust, and Meehl 1989). Clinical

judgement relates to the combination or processing of information *in the head* of a human decision-maker. By contrast, the actuarial judgement tries to eliminate the human component and conclusions are made based on empirically established relations. Assuming an antagonistic framing between the two methods, various studies in the realm of biomedical or psychiatric diagnoses find that clinical judgement is typically outperformed by simple statistical decision-rules (see also Goldberg 1970). A necessary condition for a fair comparison between the two methods is that the respective judgements use the same data. Moreover, it must be ensured by way of cross-validation that the accuracy of actuarial methods is not achieved by capitalizing on chance. The influence on the methodology of the ML approach to policy problems is evident – especially with regard to the antagonistic study design. However, one striking difference is that while actuarial methods rely on simple decision-rules, those of ML models tend to be staggeringly complex.

The second research strand are the works of Kahneman and Tversky (1974; 1979), examining how people's judgement formation about probabilities is systematically led astray by the reliance on availability heuristics, representativeness heuristics, and anchoring heuristics, and how people deviate from the axioms of Expected Utility Theory (EUT) on decisions about monetary payoffs in risky conditions. These findings, which are drawn from controlled experiments, are considered to provide evidence for the irrationality of human decision-making. The impact of Kahneman and Tversky's work can hardly be overstated. It has given rise to the field of behavioral economics, trying to revise the microfoundations of economic theorizing by integrating insights from psychology and the cognitive sciences. Closely related thereto, it animated a behavioral turn in policymaking, using insights from behavioral science as a guide for public policy interventions – most notably in the guise of Nudge theory (Thaler and Sunstein 2008) and its competitor, Boost theory (Hertwig and Grüne-Yanoff 2017).[10] Roughly, the basic idea of Nudges is to promote societal changes by way of small manipulations of the decision-architecture, thus accounting for

[10] See Malecka (2021) for a critical analysis of the kind of knowledge provided by the behavioral sciences for public policy.

people's behavioral biases. Boosts again intervene on people's decision-making competences either by improved information presentation or by providing them with simple yet effective decision-rules.

Provided that the focus of the ML approach to policy problems is on understanding and improving decision-making, it falls squarely into the paradigm of the behavioral turn to policymaking. However, compared to the works of Kahneman and Tversky and the subsequent research in psychology or behavioral economics, one important methodological difference is that there is a shift from controlled experiments towards the analysis of large-scale sets of retrospective data. This methodological shift poses many advantages. Instead of investigating the decision-making of a limited number of research subjects under stylized conditions, it has now become possible to study thousands of actual decisions. However, relying on retrospective data also has some caveats. Most importantly, the observed outcomes might have been affected by many confounding factors, not captured by the data. These confounders must be controlled for by conducting various empirical tests. In a similar vein, it is still unclear how the insights generated by the ML approach to policy problems can be translated into policy interventions. To get a better grip on these issues, we move on to a detailed discussion of the study by Mullainathan and Obermeyer (2022), trying to counteract inefficiencies in medical testing with ML models.[11]

4. Predicting Inefficiencies in Healthcare

It is commonly assumed that resources in the healthcare system are spent inefficiently – culminating in overdiagnosis and deteriorating patient outcomes. This problem is particularly pressing in medical testing, where it can be attributed to two causes. On the one hand, medical imaging techniques enable early screenings, leading to diagnoses that do not benefit patients because the diagnosed condition is not a harmful disease (Rogers and Mintzker 2016). On the other hand, there is the issue of moral hazard, diminishing the incentives to economize on treatment. Moral hazard

[11] We chose the study since we deem its methodology to be representative for a lot of research that takes place at the intersection of ML and public health. For a likeminded study, see Hastings et al. (2020).

is intimately linked to informational asymmetries between patients, physicians, and insurance providers, incentivizing physicians to perform expensive medical tests on patients (Hausman 2021). Beginning with a study by Abaluck et al. (2016), the view is gaining traction that the best way to tackle moral hazard involves examining physicians' testing intensity conditional on observable risk factors in patients. Understanding why and how physicians allocate resources inefficiently is key to identifying adequate interventions for ameliorating moral hazard.

4.1 Case Study: On Testing Patients for Heart Attack

The objective of the study by Mullainathan and Obermeyer (2022) is to investigate how physicians diagnose heart attack. Testing for heart attack involves a blockage in the coronary arteries, which is a costly and invasive procedure. Even though diagnosing heart attack is a standard procedure for physicians in the emergency room, there are many ambiguities in the diagnosis, especially since many benign conditions share symptoms with heart attack. Diagnosing heart attack requires integrating a diverse set of data. For this reason, it can be best understood as a prediction problem.

To study how physicians predict the risk of heart attack, Mullainathan and Obermeyer use a data-corpus of 250,000 emergency visits at a large academic hospital, registering patients' health records, tests given, resulting treatment, and patient outcomes. They then train an ML model to predict the outcome of testing, with the information available to physicians at the time of testing. The data are split into a training and a hold-out set. Importantly, the hold-out set is not used to benchmark physicians' predictions against the ML model but to predict patient subgroups, where physicians might have erred. To validate whether physicians indeed made errors, Mullainathan and Obermeyer consider the patients' health outcomes, assessing whether they show any major adverse cardiac events within 30 days of their visit. The testing efficiency is estimated by way of quality-adjusted life-year (QALY).[12] One reason for looking at patients' outcomes is that there is an informational asymmetry between the physician and the ML model. The model can assess electronic health

[12] See Herlitz (2018) for how QALYs are calculated.

records but is unable to capture additional information – e.g., self-reports by the patients, how they look, or results from ECGs and X-rays. Overall, the study finds that physicians overtest low-risk patients, while failing to test many apparently high-risk patients (pp. 680-1).

Mullainathan and Obermeyer formulate two hypotheses why physicians under-/overtest: The first one is that – when compared to the ML model – physicians rely on an overly simple risk-prediction model and overregularize by giving too much weight to salient features in turn. The second hypothesis is tied to the fact that health care models have fueled moral hazard by paying for tests, rather than for outcomes. An important finding of the study is that the implementation of existing moral hazard amelioration strategies, such as incentivizing low-testing regimes, can have perverse effects: it will lead to a decrease in overtesting among low-risk patients, while at the same time increasing undertesting of high-risk patients.

To rule out that high-risk patients were not tested for other reasons (e.g., they might have been too frail), the average age of the untested population and the fact whether they at least received an ECG were checked for (p. 701). Yet Mullainathan and Obermeyer acknowledge that the evidence provided for undertesting in physicians is only indirect. A potential confounding factor is that patients are first seen by nurses at the triage desk, before being examined by a physician. The involvement of the nurses in turn can influence the downstream decisions made by the physicians. For this purpose, a *natural experiment* is performed, in which the physicians' predictions are plotted at different shifts, assessing whether the error-rate proves to be stable across different triage teams (pp. 704-6).

To better understand why physicians are prone to making testing errors, it was analyzed how models for predicting physicians' testing decisions deviate from models for predicting the actual risk. The starting point here is that physicians' decision-making is boundedly rational: They are unable to process all the information available and compensate for their cognitive limitations by way of biases and heuristics (see also Wheeler 2020). This is contrasted by the complexity of the

optimal ML model that includes more than 16,000 variables. To investigate physicians' decision-making, Mullainathan and Obermeyer fit models at varying levels of complexity.[13] By testing the different models against the hold-out data, they found that the model that best predicts physician choices uses 49 variables, while the optimal risk-prediction model uses 224.

Note that the aim here is solely to shed light on the level of complexity in physicians' decision-making but not to approximate their *actual* decision-rule. In this respect, Mullainathan and Obermeyer make no assumptions regarding the properties of the models during training (p. 711). However, as a next step, they examine whether certain salient variables are overweighted by physicians – e.g., chest pain, age, and sex. This is meant to account for potential biases in physicians' testing-decisions. Here, the strategy involves training a risk-prediction model with a subset of variables, denoting salient symptoms. Despite the restrictions concerning the input data, the model has been trained exactly the same way as the original risk-prediction model, against which it was benchmarked in turn.

Again, the upshot is that the risk from symptoms is particularly predictive of physicians' testing, suggesting that, as a category, these salient symptoms are overweighted. In addition, the model's predictive accuracy has been tested for patients lacking said symptoms (pp. 716-720). The importance of empirical tests and models of varying complexity as a means to understanding ML models is an intriguing methodological insight from the study. Contrary to a suggestion that is commonly put forward within the philosophical debate on ML in science and, thus, somewhat surprisingly, methods that explain the model's behavior post-hoc through summary statistics or by detecting counterfactual dependencies virtually play no role for the study. The iterative process, beginning with theoretical assumptions on the researchers' side, followed by increasingly nuanced empirical tests, might be best captured by Sullivan's (2022) notion of 'link uncertainty' according

[13] Where model complexity is measured in terms of the variables included in a given model.

to which an understanding of and with an ML model is achieved by connecting the model behavior to background theory.

One problem of the study is that it relies on data from a single hospital, which is why there are concerns regarding the transferability of the ML model to novel settings. The testing behavior of the respective physicians may not be representative of physicians across other hospitals, which is why the ML model might overfit to the idiosyncrasies of that particular cohort. As a means to ensure its external validity, the ML model is tested against a nationally representative dataset of Medicare fee-for-service patients. Here, too, it was confirmed that physicians undertest high-risk patients while overtesting low-risk patients. However, as Mullainathan and Obermeyer (2022) point out, the dataset has significant limitations in that it is based on insurance claims, rather than EHR data and includes only few patient information. The heterogeneity and lack of standardization in medical data highlights the obstacles in validating ML models for healthcare purposes, especially since only very few external benchmark datasets exist. This may particularly affect the model's performance for marginalized social groups – traditionally underrepresented in medical data.[14]

4.2 From Predictions to Interventions

It is obvious that Mullainathan and Obermeyer (2022) should be commended for many thoughtful methodological choices. Even still, what lessons can be drawn from the perspective of a policymaker, interested in diminishing moral hazard in medical testing? Here, the risk-prediction model by Mullainathan and Obermeyer turns out to be a mixed bag. Public health interventions can take many forms, with the most obvious candidate being the release of the policy that changes the incentive structure for physicians. For instance, one might incentivize testing by other means than via changes to the reimbursement schemes or by way of capacity constraints. However, the simultaneous presence of under- and overtesting suggests that such a policy is predestined to miss its mark, since it is insufficiently nuanced. Based on another empirical test, in which the behavior

[14] See Seyyed-Kalantari et al. (2021) for a good overview about how biases in medical data culminate in unfair treatment for marginalized social groups.

of lower-testing staff is compared to higher-testing staff, Mullainathan and Obermeyer find that a reduction in tests will equally affect low-risk and high-risk-patients. Thus, even though the policy would decrease overtesting, a low-testing regime could create perverse incentives in that it leads to worse outcomes for high-risk patients. The data also indicates that the accuracy does not increase if physicians were to test less (pp. 720-722).

Similar issues are likely to arise regarding other policy proposals. Hausman (2021) recently argued that the best strategy to diminish moral hazard is to intervene on non-monetary incentives.[15] For example, by imposing non-monetary costs, so-called *ordeals*, on different agents in the healthcare-system, certain choices – e.g., abundant testing – become less attractive.[16] Imposing ordeals increases the likelihood that physicians will conduct due diligence in their medical testing. However, just as in the case of changes to monetary incentives, the result might either be an increase in tests – benefitting undertested high-risk patients, while harming low-risk patients – or the other way around. After all, neither monetary nor non-monetary incentives manage to address the cognitive limitations of physicians that ground inefficiencies in medical testing. As an upshot, one benefit of the risk-prediction model is that it enables understanding why certain policies are not adequate for a given purpose.

That being said, if the aim is to identify appropriate strategies for diminishing moral hazard in medical testing, then the risk-prediction model's level of granularity turns out to be brittle. This is due to instrumental and – more speculatively – metaphysical reasons. Concerning the former, the underlying issue is that although empirical tests facilitate a high-level understanding of the risk-prediction model, the actual decision-rule is still left obscure.[17] It is therefore difficult to derive an epistemically accessible policy that accounts for optimal testing-strategies from the model. Turning

[15] However, note that Hausman's primary aim is overcoming health inequalities that result from moral hazard.
[16] Any measure that does not impose financial costs, but rather non-monetary burdens such as "waiting or filling out forms" (Hausman 2021, p. 29) can be considered an example for such ordeals.
[17] See Creel (2020) and Boge (2022) for detailed analyses of the opacity problem in ML.

to the latter, the challenge is to identify a policy that fits to complicated natural phenomena yet, for practical purposes, must be simple (see also Mullainathan and Obermeyer 2022, p. 715).

An alternative to policies are cognitive interventions. Unlike policies that aim at solving societal problems by inducing systemic changes, cognitive interventions shift the locus of attention to improving the decision-making capacity of individual physicians.[18] Nudges or Boosts are the natural candidates here (Thaler and Sunstein 2008; Hertwig and Grüne-Yanoff 2017). As the study by Mullainathan and Obermeyer (2022) has shown, one source of error in medical testing is that physicians overweight certain salient variables. Although Nudges and Boosts are driven by different research programs, in both cases, the strategy for overcoming inefficiencies in testing might be to counteract the salience of variables by intervening on the information presentation.[19]

Nevertheless, the prospects of using the risk-prediction model by Mullainathan and Obermeyer as a guide for designing Nudges/Boosts are limited. To begin with, their study only highlights *that* a subset of variables/symptoms gets overweighted by physicians, thereby biasing testing decisions. However, this does not explain *why* physicians attach greater importance to particular variables in their decisions. More importantly, if we accept Mullainathan's and Obermeyer's premises, then Nudges and Boosts again prove to be a too simple solution for tackling a complicated phenomenon. However, this does not preclude that Nudges and Boosts can be useful in improving clinical decision-making. Rather, they are unlikely to simultaneously mitigate undertesting and overtesting. Of course, it could also be that physicians' overtesting and undertesting can be tackled by combining systemic interventions and Nudges or Boosts. However, the feasibility of such a mixed strategy is an empirical matter that can hardly be settled in this chapter.

[18] The distinction between systemic and individual interventions is based on work by Chater and Loewenstein (2022). They critically argue that the individualistic framing, typically inspired by the behavioral sciences, distracts from concerted systemic efforts by way of regulation or taxation. Those are deemed more efficient for overcoming issues such as obesity, addiction, or climate change.

[19] See Last et al. (2021) for a systematic review of physician-directed Nudges in healthcare. For Boosts, see also Marewski and Gigerenzer (2012).

This leaves policymakers with yet another cognitive intervention, namely, using the risk-prediction model as a decision-support tool for physicians. However, there are some crucial differences between the way that the algorithmic decision-support tool works as a cognitive intervention, when compared to Nudges and Boosts. Roughly, the latter seek to fix the cognitive limitations of the physician, while respecting their decisional authority – at least if one subscribes to the view that Nudges do not necessarily bypass the physicians' reasoning processes (Levy 2019). In contrast, since the decision-support tool is vastly superior to the physician at predicting patients' risk of heart attack, the physician's role is basically to be deferential. The sole reason for not straightforwardly replacing physicians with risk-prediction models is that there are certain kinds of information that are elusive to the risk-prediction model, potentially confounding the risk estimate. Provided that a patient has a red face or reports that they are feeling pain in the chest area, then this information can be used to override the risk-prediction model (Mullainathan and Obermeyer 2022, p. 723).

And yet, the epistemic rules about when to override algorithmic predictions are not well-understood, especially since the information that risk-prediction models and physicians use is asymmetric in a two-fold way. Both will use a shared pool of information – e.g., a subset of the EHR data. The risk-prediction model, however, will use the comprehensive information of the EHR data, while the physician may consider information from additional diagnostic modalities (see also Ludwig and Mullainathan 2021, p. 86). In light of this, how to sort out disagreements? Intriguingly, these informational asymmetries pose a challenge for epistemological theories of disagreement – most pertinently the Equal Weight View – that so far are predominantly concerned with cases in which the disagreement between peers is tied to the same body of evidence (Christensen 2007; see also Grote forthcoming). Empirical investigations of the interplay between physicians and decision-support tools have shown that especially novices are prone to blindly following algorithmic recommendations (Gaube et al. 2021; Tschandl et al. 2020; see also Genin and Grote 2021).

We will not engage with the question how this interplay can be improved in this chapter; be it via careful epistemological theorizing or, more pragmatically, by way of best-practice models and regulations. For present purposes, what is at stake is that when purely predictive ML models are used for solving policy problems, then this is, in all likelihood, going to result in a novel kind of cognitive intervention. In addition, public health is consequentialist in its basic construction: Provided that the involvement of risk-prediction models promotes the health of populations – understood as an aggregate of individual health levels (Munthe 2008, p. 40) – there is nothing ethically troublesome *per se* in the automation of medical testing. That being said, if the aim is to identify systemic interventions for fixing moral hazard, causal inference techniques may prove to be indispensable after all. For example, a pronounced account of the conditional probabilities between the effect variables and the target variable potentially facilitates tailoring an intervention that manages to simultaneously address the issues of undertesting and overtesting. With that in mind, we hope that this chapter provides convincing arguments for why even in the absence of causal inference techniques, predictive models should be explored as a tool for guiding public health interventions.

5. Conclusion

The chapter's leitmotif was the question why the enthusiasm for applying ML models in clinical medicine has not been echoed in the realm of public health. The answer is rooted in skepticism against purely predictive models, who are claimed to be neither sufficiently robust nor sufficiently informative to guide public health interventions. Against the widely held view that these deficits should be accounted for by blending ML models with causal inference techniques, we have argued that even purely predictive models can be a useful guide for public health interventions. To this end, we considered a novel strand of research that uses ML models for solving prediction policy problems. This research strand, again, can be seen as a part of the behavioral turn in policymaking. However, there are clear discontinuities regarding the methodology and the resulting kinds of interventions, when compared to existing policy-approaches grounded in the behavioral sciences

such as Nudges and Boosts. Based on a case-study from health economics, we discussed in detail how predictive models can foster understanding of inefficiencies in medical testing, while also providing tools for improving public health outcomes. However, the utility of ML models is restricted in a two-fold way: First, they can only be meaningfully used for narrowly confined routine tasks. Second, they are most efficient when the adequate intervention is already known in advance. Alternatively, the ML model itself may turn into a cognitive intervention, providing decision-support for policymakers.

By discussing studies from health economics, we also contribute to the understanding of the specific methodological challenges of ML-driven science outside of *traditional* success areas – e.g., structural biology, climate science, or medical imaging. In particular, we have highlighted the relevance of empirical tests and comparator models for overcoming the opacity of ML models. Our account has also wider-ranging implications for the (philosophical) debate on evidence-based medicine/policymaking. Given their empiricist foundations, the epistemic value of predictive models is largely ignored by existing frameworks. Instead, they are preoccupied over the caveats of RCTs for establishing causal claims, or the complicated liaison between causal and mechanistic evidence (Deaton and Cartwright 2018; Russo and Williamson 2007; Marchionni and Reijula 2019; Shan and Williamson 2021). In light of the increasing attention that ML models receive in medicine and the social sciences, it is about time to diversify perspectives, reflecting on what constitutes good predictive models for policymaking, and how the relevant evidence can be meaningfully combined with other kinds of evidence (see also Broadbent 2013; Boumans 2019).

Acknowledgements

TG is supported by the Deutsche Forschungsgemeinschaft (BE5601/4-1; Cluster of Excellence "Machine Learning—New Perspectives for Science", EXC 2064, project number 390727645). TG also acknowledges support by the project "Certification and Foundations of Safe Machine Learning Systems in Healthcare" funded by the Carl Zeiss Foundation.

OB is supported by an ETH Zürich Postdoctoral Fellowship.

Competing Interests

The authors have no conflicts of interest to declare that are relevant to the content of this chapter.

References

Abaluck, J., Agha, A., Kabrhel, C., Raja, A., Venkatesh, A. (2016): The Determinants of Productivity in Medical Testing: Intensity and Allocation of Care. *American Economic Review*, 106(12):3730-3764.

Athey, S. (2017): Beyond Prediction: Using Big Data for Policy Problems. *Science*:355(6324):483-485.

Boge, F. J. (2022). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*, *32*(1), 43-75.

Boumans, M. (2019): Simulation and Economic Methodology. In: Dolfsma, W., Wade Hands, D., McMaster, R. (eds.): *History, Methodology and Identity for a 21st Century Social Economics*. Routledge, London, pp. 41-50.

Broadbent, A. (2013): *Philosophy of Epidemiology*. Palgrave Macmillan, New York.

Broadbent, A., Grote, T. (2022): Can Robots Do Epidemiology? Machine Learning, Causal Inference, and Predicting the Outcomes of Public Health Interventions. *Philosophy & Technology*, 35(1):1-22.

Buchholz, O. (2023a): A Means-End Account of Explainable Artificial Intelligence. *Synthese*, 202:33.

Buchholz, O. (2023b): The Deep Neural Network Approach to the Reference Class Problem. *Synthese*, 203:111.

Buchholz, O., Grote, T. (2023): Predicting and Explaining with Machine Learning Models: Social Science as a Touchstone. *Studies in History and Philosophy of Science*, 102:60-69.

Chater, N., Loewenstein, G. (2022): The i-frame and the s-frame: How Focusing on the Individual-level Solutions Has Led Behavioral Public Policy Astray. *SSRN Preprint 4046264*, URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4046264

Christensen, D. (2007): Epistemology of Disagreement: The Good News. *The Philosophical Review*, 116(2):187-217.

Creel, K. A. (2020): Transparency in Complex Computational Systems. *Philosophy of Science*, 87(4):568-589.

Cui, P., Athey, S. (2022): Stable Learning Establishes Some Common Ground between Causal Inference and Machine Learning. *Nature Machine Intelligence*, 4(2):110-115.

Dawes, R. M., Faust, D., Meehl, P. E. (1989): Clinical versus Actuarial Judgment. *Science*, 243(4899):1668-1674.

Deaton, A., Cartwright, N. (2018): Understanding and Misunderstanding Randomized Controlled Trials. *Social Science and Medicine*, 210:2-21.

Díaz-Lozano, M., Guijo-Rubio, D., Gutiérrez, P. A., Gómez-Orellana, A. M., Túñez, I., Ortigosa-Moreno, L., ..., Hervás-Martínez, C. (2022): COVID-19 Contagion Forecasting Framework Based on Curve Decomposition and Evolutionary Artificial Neural Networks: A Case Study in Andalusia, Spain. *Expert Systems with Applications*, 117977.

Einav, L., Finkelstein, A., Mullainathan, S., Obermeyer, Z. (2018): Predictive Modeling of US Health Care Spending in Late Life. *Science*, 360(6396):1462-1465.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., Thrun, S. (2017). Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. *Nature*, 542(7639):115-118.

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., ..., Saria, S. (2021): The Clinician and Dataset Shift in Artificial Intelligence. *The New England Journal of Medicine*, 385(3):283.

Freiesleben, T., & Grote, T. (2023). Beyond generalization: a theory of robustness in machine learning. *Synthese*, *202*(4), 109.

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., ..., Ghassemi, M. (2021): Do as AI Say: Susceptibility in Deployment of Clinical Decision-aids. *NPJ Digital Medicine*, 4(1):1-8.

Gelman, A. (2010): Causality and Statistical Learning. URL: https://arxiv.org/ftp/arxiv/papers/1003/1003.2619.pdf.

Genin, K., Grote, T. (2021): Randomized Controlled Trials in Medical AI: A Methodological Critique. *Philosophy of Medicine*, 2(1), DOI: 10.5195/pom.2021.27.

Goldberg, A. L. (1970): Man Versus Model of Man: A Rationale, Plus Some Evidence, for a Method of Improving on Clinical Inferences. *Psychological Bulletin*, 73:422-432.

Grote, T. (forthcoming): Machine Learning in Healthcare and the Methodological Priority of Epistemology over Ethics. *Inquiry*, DOI: 10.1080/0020174X.2024.2312207.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., ..., Webster, D. R. (2016): Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22):2402-2410.

Hardt, M., Recht, B. (2022): *Patterns, Predictions, and Actions. A Story About Machine Learning*. Princeton University Press, Princeton.

Hastings, J. S., Howison, M., & Inman, S. E. (2020). Predicting high-risk opioid prescriptions before they are given. *Proceedings of the National Academy of Sciences*, *117*(4), 1917-1923.

Hausman, D. M. (2021): Ordeals, Inequalities, Moral Hazard and Non-monetary Incentives in Health Care. *Economics & Philosophy*, 37(1):23-36.

Hernán, M. A., Hsu, J., Healy, B. (2019): A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *Chance*, 32(1):42-49.

Hertwig, R., Grüne-Yanoff, T. (2017): Nudging and Boosting: Steering or Empowering Good Decisions. *Perspectives on Psychological Science*, 12(6):973-986.

Herlitz, A. (2018): Against Lifetime QALY Prioritarianism. *Journal of Medical Ethics*, 44(2):109-113.

Hinton, G., Vinyals, O., Dean, J. (2015): Distilling the Knowledge in a Neural Network. *arXiv preprint*, URL: https://arxiv.org/abs/1503.02531.

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., ..., Yarkoni, T. (2021): Integrating Explanation and Prediction in Computational Social Science. *Nature*, 595(7866):181-188.

Hyland, S. L., Faltys, M., Hüser, M., Lyu, X., Gumbsch, T., Esteban, C., ..., Merz, T. M. (2020): Early Prediction of Circulatory Failure in the Intensive Care Unit Using Machine Learning. *Nature Medicine*, 26(3):364-373.

Kahneman, D., Tversky, A. (1979): Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2):263-292.

Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z. (2015): Prediction Policy Problems. *American Economic Review*, 105(5):491-95.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S. (2018): Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133(1):237-293.

Last, B. S., Buttenheim, A. M., Timon, C. E., Mitra, N., Beidas, R. S. (2021): Systematic Review of Clinician-directed Nudges in Healthcare Contexts. *BMJ Open*, 11(7):e048801.

Lazer, D., Kennedy, R., King, G., Vespignani, A. (2014): The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343(6176):1203-1205.

Levy, N. (2019): Nudge, Nudge, Wink, Wink: Nudging Is Giving Reasons. *Ergo*, 6:10.
Ludwig, J., Mullainathan, S. (2021): Fragile Algorithms and Fallible Decision-makers: Lessons from the Justice System. *Journal of Economic Perspectives*, 35(4):71-96.
Małecka, M. (2021): Knowledge, Behaviour, and Policy: Questioning the Epistemic Presuppositions of Applying Behavioural Science in Public Policymaking. *Synthese*, 199(1):5311-5338.
Marchionni, C., Reijula, S. (2019): What Is Mechanistic Evidence, and Why Do We Need It For Evidence-based Policy? *Studies in History and Philosophy of Science Part A*, 73:54-63.
Marewski, J. N., Gigerenzer, G. (2012). Heuristic Decision Making in Medicine. *Dialogues in Clinical Neuroscience*, 14(1):77-89.
Maslow, A. H. (1966). *The psychology of science: A reconnaissance*. Harper & Row.
Mullainathan, S., Obermeyer, Z. (2022): Diagnosing Physician Error: A Machine Learning Approach to Low-value Health Care. *The Quarterly Journal of Economics*, 137(2):679-727.
Mullainathan, S., Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2):87-106.
Munthe, C. (2008): The Goals of Public Health: An Integrated, Multidimensional Model. *Public Health Ethics*, 1(1):39-52.
Northcott, R. (2017): When Are Purely Predictive Models Best? *Disputatio*, 9(47):631-656.
Northcott, R. (2020): Big Data and Prediction: Four Case Studies. *Studies in History and Philosophy of Science Part A*, 81: 96-104.
Northcott, R. (2022): Reflexivity and Fragility. *European Journal for Philosophy of Science*, 12(3):1-14.
Pearl, J., Mackenzie, D. (2019). *The Book of Why: The New Science of Cause and Effect*. Penguin Books, London.
Perdomo, J., Zrnic, T., Mendler-Dünner, C., Hardt, M. (2020): Performative Prediction. In *Proceedings of the International Conference on Machine Learning*, pp. 7599-7609
Rogers, W. A., Mintzker, Y. (2016): Getting Clearer on Overdiagnosis. *Journal of Evaluation in Clinical Practice*, 22(4):580-587.
Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
Russo, F., Williamson, J. (2007): Interpreting Causality in the Health Sciences. *International Studies in the Philosophy of Science*, 21(2):157-170.
Salganik, Matthew J., et al.(2020) "Measuring the predictability of life outcomes with a scientific mass collaboration." *Proceedings of the National Academy of Sciences* 117(15): 8398-8403.
Schölkopf, B., Locatello, F., Bauer, S., Ke, N., Kalchbrenner, N., Goyal, A., Bengio, Y. (2021): Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5):612-634.
Seyyed-Kalantari, L., Zhang, H., McDermott, M., Chen, I. Y., Ghassemi, M. (2021): Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in Under-served Patient Populations. *Nature Medicine*, 27(12):2176-2182.
Shan, Y., & Williamson, J. (2021). Applying Evidential Pluralism to the social sciences. *European Journal for Philosophy of Science*, *11*(4), 1-27.
Sullivan, E. (2022): Understanding from Machine Learning Models. *The British Journal for the Philosophy of Science*, 73(1):109-133.
Syrowatka, A., Kuznetsova, M., Alsubai, A., Beckman, A. L., Bain, P. A., Craig, K. J. T., ..., Bates, D. W. (2021): Leveraging Artificial Intelligence for Pandemic Preparedness and Response: A Scoping Review to Identify Key Use Cases. *NPJ Digital Medicine*, 4(1):1-14.
Thaler, R., Sunstein, C. (2008): *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press, London.

Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., ..., Kittler, H. (2020): Human–computer Collaboration for Skin Cancer Recognition. *Nature Medicine*, 26(8):1229-1234.

Tversky, A., Kahneman, D. (1974): Judgment under Uncertainty: Heuristics and Biases: Biases in Judgments Reveal Some Heuristics of Thinking under Uncertainty. *Science*, 185(4157):1124-1131.

Verweij, M., Dawson, A. (2007): The Meaning of 'Public' in 'Public Health'. In Dawson, A, Verweij, M. (eds.): *Ethics, Prevention, and Public Health*. Oxford University Press, New York, pp. 13-29.

Wheeler, G. (2020): Bounded Rationality. In: Zalta, E. N. (ed.): *The Stanford Encyclopedia of Philosophy*, URL: https://plato.stanford.edu/archives/fall2020/entries/bounded-rationality/.

About the authors

Thomas Grote is a research fellow at the Cluster of Excellence: "Machine Learning: New Perspectives for Science" at the University of Tübingen. He is also Co-PI in a project on certification and safety of ML models in healthcare, funded by the Carl-Zeiss Stiftung.

Oliver Buchholz is a postdoctoral research fellow at the Chair of Bioethics at ETH Zurich and an associate member of the Interchange Forum for Reflecting on Intelligent Systems at University of Stuttgart.