Article

Philosophy of the Social Sciences
2023, Vol. 53(5) 371–395

DOI: 10.1177/00483931231172455
journals.sagepub.com/home/pos
Sage

Managing Performative Models

Donal Khosrowi[1]

Abstract

Scientific models can be performative: they can causally affect the phenomena they are intended to represent. The existing literature offers two responses. The *appraisal view* emphasizes that performativity can sometimes be a good-making model attribute, e.g., when predictions steer the public's behavior in desirable ways. The *mitigation view* seeks to endogenize agents' behavioral response to model-issued forecasts to get rid of performativity instead. This paper argues that neither approach is fully compelling: the appraisal view encounters severe concerns about moral values illegitimately encroaching on how modelers construct and use models, while the mitigation view fails to acknowledge that endogenization is itself a choice that involves substantive value-judgments relating to the desirability of certain social outcomes.

Keywords

performativity, models, policy advice, endogenization, values in science

1. Introduction

Scientific models can be performative: in addition to serving various epistemic purposes, they can also causally affect phenomena, such as when agents' behaviors change in response to model predictions. In recent years, philosophers have made substantial progress in delineating different forms of performativity and characterizing the problems they can pose, such as when

[1] Leibniz Universität Hannover, Germany

Corresponding Author:
Donal Khosrowi, Leibniz University Hannover, Lange Laube 6, Hannover 30159, Germany.
Email: donal.khosrowi@philos.uni-hannover.de

the forecasts researchers derive from models are self-defeating and compromise models' epistemic functioning (Avery et al. 2020b; Godman and Marchionni 2022; Jiménez-Buedo 2021; Northcott 2022; Tee 2019; van Basshuysen 2022; van Basshuysen et al. 2021; Vergara-Fernández, Heilmann, and Szymanowska 2023; Winsberg and Harvard 2022).

The existing literature offers two broad types of response to model performativity. First, to maintain models' predictive performance one can *endogenize*, i.e., explicitly model how agents will respond to a prediction and accommodate this response in the predictions made. This *mitigation* approach has been pursued by social scientists as early as the 1950s (Grunberg and Modigliani 1954; Simon 1954) and currently enjoys renewed interest (Avery et al. 2020b; Perdomo et al. 2021). A second approach was recently outlined by Philippe van Basshuysen, Lucie White, Mathias Frisch, and myself in the context of epidemiological models informing policy responses to the SARS-CoV-2 pandemic (van Basshuysen et al. 2021). There, we argued that performativity can sometimes be understood as a good-making model attribute, e.g., when predictions derived from models, such as that critical care demand will exceed capacity, steer the public's behavior in desirable directions. Our *appraisal* approach hence understands models as tools that have both epistemic and performative capabilities, both of which should be considered in model evaluation, and permits (some forms of) performativity to count as a good-making feature of models (see Vergara-Fernández, Heilmann, and Szymanowska 2023 for related proposals).

In this paper, I argue that neither approach is fully compelling.[1] The appraisal approach recognizes that performative models may have good-making performative features, but, as we stress in van Basshuysen et al. (2021), struggles with providing guidance for adjudicating models' epistemic and performative roles when they are in tension. Specifically, while it might sometimes seem appropriate to appraise models post-hoc for having made performative contributions (e.g., helping agents manage their response to a new wave of SARS-CoV-2 infections), performativity should not figure as a criterion in model construction since doing so can incentivize unacceptable value influences to encroach on the construction and use of models and may threaten the epistemic integrity of model-based science (cf. Winsberg and Harvard 2022).

The mitigation approach, by contrast, does not seem to get into such murky waters. It maintains that performativity is a phenomenon that can be kept in check by endogenizing individuals' behavioral response to model outputs. However, I argue that in aiming to keep performativity in check, the approach disregards that (1) by "endogenizing away" agents' behavioral response to

[1] Note that the arguments developed here reflect my own views, but not necessarily those of my co-authors in van Basshuysen et al. (2021).

align forecasts with actual behaviors, it neglects the potential real-world pragmatic benefits that performative models can harbor, and (2) because endogenizing behavioral response can prevent such benefits from obtaining, endogenization is itself a choice that involves substantive value-judgments. There is hence no value-neutral stance when deciding whether to let model outputs influence behaviors or to prevent such effects from obtaining.

With neither option able to keep important value-related concerns at bay, I offer some constructive proposals for managing performativity, i.e., acknowledging models' performative contributions, while ensuring that their epistemic integrity remains uncompromised. I especially focus on carving out clearer principles to help keep value-influences from illegitimately meddling with the production and use of models to inform policy. Several decision points concerning model construction and use must be kept independent of researchers' views regarding the desirability of potential performative effects. What is more, while decision-makers may legitimately make value-laden choices about how to interpret model outputs, how to use them in decision-making, and how to communicate their decisions to the public, they must refrain from suggesting that their decisions follow straightforwardly from model outputs (e.g., claiming that they merely "follow the science"). This is to ensure that models do not carry excessive justificatory burden in grounding value-laden decisions.

The discussion is organized as follows. *Section 2* clarifies the concept of model performativity and briefly outlines the case of performative Covid models discussed in the literature. *Section 3* outlines the two major approaches for dealing with performativity, appraisal and mitigation, reconstructs and sharpens the main concerns about the appraisal strategy, and argues that mitigation is susceptible to related concerns about illegitimate value influences. *Section 4* further elaborates the central tensions arising when models have performative capacities, emphasizes important contextual features that bear on how performativity may be addressed, and proposes a general principle to mitigate the most severe value-related concerns raised by performativity. *Section 5* concludes.

2. What is Model Performativity?

Model performativity is now situated in a rich and growing conceptual forest with cognate notions tracking several related phenomena, including reactivity, reflexivity, interactivity, and others (see Buck 1963; Godman and Marchionni 2022; Henshel 1993; Jiménez-Buedo 2021; Vergara-Fernández, Heilmann, and Szymanowska 2023). I will not explore these related concepts and the interesting arguments they ground, nor offer a general account of performativity or trace its rich intellectual history (see Callon and Roth 2021; Guala 2007; Mäki 2013; MacKenzie 2006; Perdomo et al. 2021; van Basshuysen

2022). For the purposes of this paper, I understand model performativity simply according to the following broad, causal construal (cf. Buck 1963; Henshel 1993):[2]

> **(Performativity)**: A model is performative if and only if it has the capacity to causally affect an aspect of the world that it is intended to represent.

Let me add some clarifications right away. First, by *capacity* I mean an *actualized* disposition to causally affect a target. All kinds of models could have any number of *un*actualized dispositions to affect a target, but these will be bracketed here. So, a model is performative when it actively changes aspects of a target, by itself or through particular ways of using it. Second, a performative *effect* is the difference to a target system that is, causally, due to the model and its outputs. Third, in understanding model performativity, it is important to note that models are rarely performative *as such* but typically become performative only when embedded in a concrete context of use, which establishes causal connections between the model and its target (see Vergara-Fernández, Heilmann, and Szymanowska 2023). It often takes a user who does something with a model (e.g., derive and publicize a prediction) to establish such a connection. Finally, performativity can come in many different forms, but the discussion here will mostly focus on cases where a model's *predictions* affect some of the quantities to be predicted, e.g., by triggering behaviors of agents in a target that affect these quantities.[3] Several subtypes of performative model predictions are routinely distinguished (Buck 1963; Henshel 1993). For instance, *self-fulfilling* performativity obtains when a model correctly predicts a quantity X to be $X = x$, but if the model had predicted differently, or had not been used to predict X (or the causal pathways from the prediction to the target system had been disrupted, e.g., when a prediction is kept secret), the value of X would have been different, i.e., $X = x^*$. Relatedly, *self-effacing* performativity obtains when a model is used to predict $X = x$, but *because* of the model's prediction, X changes to $X = x^*$,

[2] This definition is significantly broader than many others discussed in the extant literature, which have been emphasized to track more philosophically interesting forms of performativity (cf. Boldyrev and Ushakov 2016; Guala 2007), e.g., the way in which economic theories literally bring the kinds of agents, behaviors, or phenomena they describe into existence by virtue of their normative force. The present discussion highlights that even less involved forms of performativity can nevertheless pose significant epistemic-ethical challenges that have so far remained understudied.

[3] This is different from other cases discussed in the literature where models co-shape a target through agents using the models to gain new forms of epistemic and/or practical access to or control over a target (see Guala 2007; Vergara-Fernández, Heilmann, and Szymanowska 2023).

even though X would have been $X = x$ if the model had predicted differently, or had not been used to predict X (or if the causal pathways between model and target had been disrupted) (MacKenzie 2006). In what follows, I prevalently focus on the latter type of case, which notably involves that a models' predictive performance is negatively affected by its performativity. This is not to suggest, however, that self-fulfilling performativity, where a model causally brings about states of affairs that make its predictions true, is any less philosophically interesting or epistemic-ethically challenging. There are numerous cases where, if a social outcome is realized *mainly* because a model said it would obtain, this might equally raise concerns about a models proper epistemic functioning, e.g., when an economic model predicts a financial crisis and this prediction induces market actors to bring about that crisis. Here, a model's epistemic functioning is parasitic upon its performative capacity and we may hence think that it does not adequately function as an epistemic tool.[4]

Despite some detailing, the working construal of model performativity offered here is still extremely broad (cf. Callon and Roth 2021; Perdomo et al. 2021). Virtually all models whose outputs are involved in decision-making can count as performative to some degree, e.g., when ecologists use models to inform ecosystem preservation policies and these policies are efficacious in making a difference to features of the system modeled. Or, moving away from scientific models, when a physical scale model of a building is used to make design or engineering decisions that affect how the building is eventually constructed at scale. We might hence worry that this construal of performativity is too inclusive. While I do not see serious problems with the extension of performativity being large, a first pass to focus on a narrower class of phenomena is to make a distinction of significance: model performativity of potential interest to philosophers of science is simply the subset of cases where performativity raises significant epistemic-ethical issues. So architectural models would not regularly make the cut, or really any mundane use of models to achieve unproblematic practical aims relating to changing a target.

A second, less haphazard way of detailing performativity is to explore what makes some cases seem more significant. One feature that plays an important role here is the extent to which model users and agents whose behaviors are influenced by the model are aware of a model's performativity. Conscientiously and successfully using a model to inform efforts to change a target, as such, is often unproblematic and perhaps not especially philosophically significant.[5] However, using a model for predictive purposes and remaining

[4] At least not in a narrow sense; and especially not when agents are unaware of its performativity.

[5] To be clear, while we may of course take issue with the purposes pursued by drawing on information supplied by a model, this does not imply that *the model*, or the fact *that* it is used (rather than *for* what), are epistemically or morally problematic in themselves.

unaware that a target's behavior is influenced by the very predictions that are supposed to provide epistemic access to it, can indicate that something is going wrong, epistemically, ethically, or both. Awareness of whether a model is performative comes in degrees and model builders, users, and those exposed to any performative effects may each have different degrees of such awareness. Speaking generally, it seems that the less aware relevant stakeholders are of a model's performativity, the more likely it is that some epistemically and/or ethically problematic is going on, e.g., when users of an economic model are unaware that they are not merely predicting a financial crisis, but rather facilitating it and agents remain unaware that the prediction was not inescapable but merely self-fulfilling. This is not to suggest, however, that awareness of a model's performativity makes a case unproblematic or uninteresting. For instance, it is now widely understood that models used to predict user engagement and guide what information we are exposed to on social media can be performative: they not only serve to predict what contents are most interesting but can also cement or induce interests, preferences, and behaviors (Cinelli et al. 2021). It seems that especially in cases where modelers make value-laden choices in model construction to promote or hinder certain performative effects and this ends up negatively affects agents' outcomes without them being aware of this, agents are wronged in a special kind of way: a lack of awareness or understanding often implies a lack of agency, e.g., to resist, respond to, or challenge a prediction or the dynamics triggered by it. Finally, even if relevant stakeholders are aware of a model's performativity but this performativity is unintended and/or contravenes at least some values held or goals pursued, this equally has the capacity to make a case of performativity epistemically and ethically significant, either because it points to the inadequacy of the model for the purpose at hand, or because its use has undesirable consequences for at least some stakeholders.

I expand in more detail later on additional contextual factors that can moderate whether performativity is epistemically or ethically problematic and which avenues are best for managing it. For now, let me introduce a working example, performative epidemiological models of the SARS-Cov-2 pandemic, which will help elaborate the two main strategies to manage performativity. Although the case of Covid models is, arguably, not a paradigmatic instance of performativity in social science contexts, there are several reasons for discussing it as the central case study here. First, the recent literature engaging with the epistemic-ethical aspects of model performativity has focused on this case (van Basshuysen et al. 2021; Winsberg and Harvard 2022) and the contributions offered there are best discussed and critically challenged in their original context. Second, the case of Covid models is tractable: the models used were often simple and the dynamics of how individuals may respond to model forecasts are intuitively graspable. Third, arguably, people responding to model forecasts about the trajectory of a pandemic involves causes and concepts that figure centrally in social scientific analysis, e.g., individual constrained optimization in light of tradeoffs, norms, institutions, and so on, which are familiar modes of analysis to social scientists. In virtue of this, the lessons learned from this case can also be easily transferred to other cases of interest (I point to some later on).

2.1 Performative Covid Models

In the early stages of the SARS-Cov-2 pandemic, modeling groups around the world built epidemiological models to forecast the trajectory of key pandemic variables, e.g., number of infections, deaths, and so on. In van Basshuysen et al. (2021), we argued that these forecasts recognizably shaped policy advice offered to policy makers in the US and UK: shortly after release of the infamous ICL Report 9 study (Ferguson et al. 2020), policy makers changed policies dramatically from mitigation toward aggressive suppression measures, including strict lockdowns. In addition, especially in the early stages of the pandemic, it seems likely that dramatic forecasts such as those made in Report 9, i.e., around 510k deaths in the UK and 2.2 M in the US if viral spread were left unmitigated, had direct effects on individuals' behaviors, too, e.g., people reducing contacts and self-isolating ahead of lockdown policies, or interpreting rules more strictly (Friedson et al. 2020; Sears et al. 2023). The epidemiological models in this case were performative in the sense that they causally affected some or all of the features of the target systems they represented. If people became more cautious in response to model forecasts and increased cautiousness decreased contacts and thereby infection numbers and deaths, this directly affected the quantities targeted and would hence be an archetypal case of model performativity. Importantly, performative effects such as the ones sketched in van Basshuysen et al. (2021) often imply that a model's predictive abilities are diminished: when a model is used to (publicly) predict X and X is considered undesirable by the public, individuals might respond to the prediction in ways that end up preventing X from occurring, thus undermining the predictive accuracy of the model. Performativity thus has the potential to significantly inhibit central epistemic functionings of models and may hence be considered an undesirable feature of a model.

Let me reconstruct and elaborate two competing views that deal with this problem in different ways. First, the *appraisal view*, which acknowledges that performativity can be epistemically undesirable, but maintains that performative effects can nevertheless be a good-making feature of models. Second, the *mitigation view*, which prioritizes models' epistemic functioning by getting rid of performative effects. Aiming to make progress on understanding which of these views is more plausible, I argue that while the appraisal view faces severe concerns about illegitimate value influences in model construction and use, the mitigation strategy, despite promising to do better, is vulnerable to similar concerns.

3. Responding to Model Performativity

3.1 The Appraisal View

Covid models have been widely criticized for their poor predictive performance (Avery et al. 2020a; Ioannidis, Cripps, and Tanner 2020; Winsberg, Brennan, and Surprenant 2020). Against the background of such criticisms, we argued that when assessing Covid models, we should not only look at their forecast accuracy but also at their performative contributions (van Basshuysen et al. 2021). More specifically, while there are good reasons to think that many Covid models have been predictively far from impressive, it is unclear whether this alone is enough to conclude that they were bad models, full stop. For one, we emphasize that in assessing the epistemic contributions of Covid models, we must understand a majority of their outputs as conditional forecasts[6] for counterfactual scenarios rather than as straightforward predictions of actual courses of events (see also Fuller 2021; Schroeder 2021). So, if a conditional forecast predicts millions of deaths for a scenario where no measures are taken and, in response to that, aggressive suppression measures are implemented, it should be no surprise that actual death tolls are much lower than the forecast. Due to the policy measures implemented, the relevant quantity to assess forecast accuracy is now a counterfactual quantity that cannot be observed and at best estimated (Friedman et al. 2021; Winsberg and Harvard 2022). To be sure, many Covid models have also been used to issue a range of different scenario forecasts, including some capturing scenarios that more closely resembled actual policy trajectories taken. But even when looking at these scenario forecasts, some critics maintain that forecast accuracy has been poor, with many models overestimating infection numbers and deaths (Winsberg, Brennan, and Surprenant 2020; Winsberg and Harvard 2022).

So, should we conclude that Covid models have been bad models? In van Basshuysen et al. (2021) we argued that such a conclusion would be too hasty, sketching what I call the *appraisal view*. Specifically, we maintain that we

[6] Such conditional forecasts have also been called “projections” to distinguish them from “unconditional” forecasts or “predictions” (Fuller 2021; Schroeder 2021; Winsberg and Harvard 2022). Here, I will continue to refer to conditional forecasts as “forecasts,” since any forecast is essentially conditional on some assumptions about a target and the difference rather seems to be in regard to how clearly these conditions are articulated.

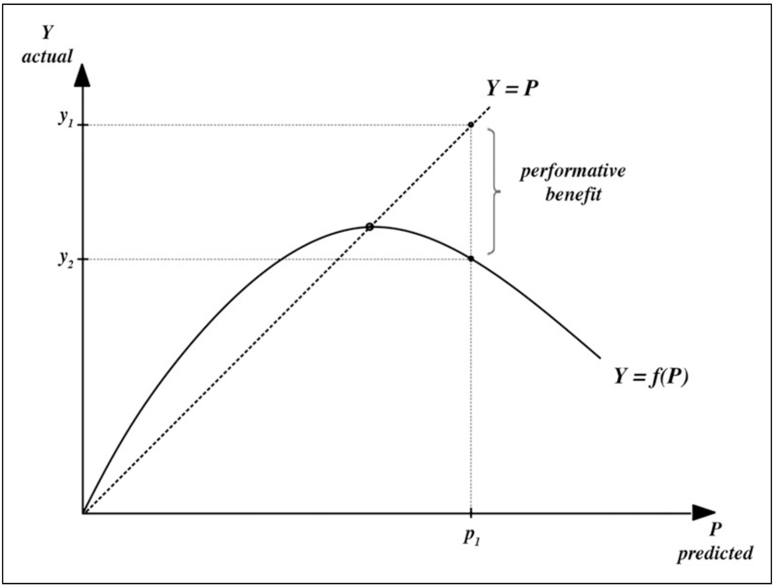

Figure 1. Beneficial performative effects.

should not only focus on forecast accuracy when evaluating Covid models, but should instead take an all-things-considered view, which takes into account their performative contributions to the achievement of (some) social goals (though possibly at the expense of others; see Winsberg and Harvard 2022). In a nutshell, the appraisal view maintains that it can be a good-making feature of Covid models that they helped individuals understand the likely trajectories of the pandemic, choosing response profiles that were (more) consistent with their preferences, and thereby contributing to lower infection numbers and death tolls. Performative effects such as these are part of what we should consider when assessing the overall goodness of models. Figure 1 captures this line of thinking in an idealized fashion.

Focusing on infection numbers as the quantity to be predicted, the P-axis plots model predictions and the *Y*-axis plots actual infection numbers. In a non-performative world with perfectly accurate models, the predictions from a model would be on the dotted 45° line, perfectly coinciding with observed values, i.e., $Y = P$. In a world with model performativity, the solid curve captures how infection numbers depend on agents' behavioral response to model predictions according to some response function $Y = f(P)$. In a nutshell, for low predicted infection numbers, individuals choose more risky behaviors, e.g., by bending lockdown rules or isolation requirements, or simply increasing contacts. In the upper right area, the response curve slope turns negative.[7] Here, high predicted numbers (P_1) of infections lead to individuals reducing contacts, thereby realizing lower numbers than initially predicted ($Y_2 < Y_1$). On the appraisal view, the difference between Y_1 and Y_2 can be understood as a benefit afforded by the model's performativity. Even though its predictions have not been accurate ($Y_1 - Y_2$), the actual outcome is, I assume here for simplicity, preferable over the alternative.[8] And since this outcome is caused, in part, by the model, achieving this outcome is a good-making feature *of* the model.

Is the appraisal view plausible? Endorsing performativity as a good-making feature may seem inappropriate for a wide range of scientific models. When a model is designed for purely epistemic purposes, it will often seem unhelpful at best and problematic at worst to think that a causal coupling between a model and target that systematically prevents a model from producing accurate forecasts can be a good-making feature. However, Covid models were constructed and used specifically for the purpose of (helping decision-makers with) inhibiting and controlling the spread of the virus, and the performative effects outlined in van Basshuysen et al. (2021) promote those same practical goals. So, broadly following an adequacy-for-purpose type view on model appraisal (Parker 2020), why should we *not* consider the achievement of these goals to be a good-making feature if models causally contributed to it?

Winsberg and Harvard (2022) argue that performative effects should never be considered good-making features of models. One of two main worries they flag is that any inaccurate forecast could always be explained as the result of performative effects and allowing performative effects to count toward a positive appraisal of a models' overall goodness would only make it easier to conjure up such ad hoc defenses. An important constraint on the appraisal view, therefore, should be that models' *counterfactual* forecasts (sans performativity) must be approximately accurate:[9] if a model forecasts high infection numbers (Y_1),[10] but much lower numbers are eventually observed

[7] Though note that the shape of the curve could in principle look different (e.g., monotonically decreasing but strictly positive in slope), and might indeed be vastly different in other cases, e.g., slope >1 throughout.

[8] The assumption that fewer infections are preferable over more is, of course, a highly contentious one since it disregards the costs associated with this putative benefit, e.g., infractions of civil liberties due to lockdowns and other restrictions, psychological and economic costs, etc. The assumption is hence made here only for the sake of the argument, and readers are asked to imagine a scenario where the cost-benefit profile of moving from Y_1 to Y_2 is indeed positive.

[9] I.e., accurate within the envelope of the models' predictive abilities, and assuming that modelers have made sincere attempts to promote these abilities.

[10] for a scenario most closely resembling the policy trajectory taken.

(Y_2), it must be true that the difference between Y_1 and Y_2 is prevalently due to performative effects and not simply due to the model getting Y_1 wrong while also failing to anticipate how individuals' response to Y_1 would move us to Y_2. Even under this narrower constraint, however, Winsberg and Harvard emphasize that we should never celebrate inaccurate, pessimistic projections as good-making features of models—even if they contributed to moving behavioral response in prima facie beneficial ways (2022, 5). Specifically, they argue that the main purpose of Covid models was to aid decision-makers[11] with understanding the cost-benefit profile of potential courses of actions. So, if Covid models systematically overestimated numbers (i.e., they did not just fail to predict Y_2, but got Y_1 wrong as well), then those models would distort the cost-benefit profile of available actions by unduly exaggerating some of the costs and benefits involved (e.g., how many lives could be saved). Of course, due to lack of reliable epistemic access to the relevant counterfactuals, telling whether models get their counterfactuals approximately right is extremely difficult in practice. But as the worries flagged by Winsberg and Harvard make clear, model appraisal must at least involve sincere attempts to assess whether this is so. So, when modelers explain away prima facie predictive failures and defend the goodness of their model by unsubstantiated blanket appeals to performativity, these defenses need to be scrutinized, and we may reasonably require modelers to offer compelling grounds to think that their models (1) *did* get their counterfactuals approximately right and (2) performativity is *really* what explains the differences between forecasts and actual, observed outcomes.

Even if such efforts were successful, however, there is a second important problem with the appraisal view that we anticipated in van Basshuysen et al. (2021) and that Winsberg and Harvard further discuss. To appreciate this problem, let me cast the appraisal view in somewhat clearer outlines:

> **Appraisal:** The overall goodness of a model (e.g., in terms of a wide understanding of *adequacy-for-purpose*; see Parker 2020; van Basshuysen 2022), is a function of (1) whether a model properly performs its *epistemic* functions, e.g., issuing accurate predictions, providing adequate explanations or facilitating understanding of a phenomenon, and (2) whether a model contributes to the achievement of the *practical* purposes for which the model was constructed, including by causally affecting desired kinds of change in a target system (see Tee 2019).

As we emphasize in van Basshuysen et al. (2021, 123), the second condition can be met in two significantly different ways, giving rise to two

[11] Understood widely here to include not only policy-makers but citizens, too.

quite different renditions of the appraisal view. To see this, let us assume that there is a set of moral and political values V, shared by stakeholders in a population. Assume model M is built to promote a set of practical purposes P (e.g., managing a pandemic) that cohere with V in a target T. One rendition of the appraisal view is *evaluative*. On this rendition, a model M is, other things being equal, a better model if it made larger differences to the achievement of P. As the tense suggests, this rendition is *backward-looking*: given a model that has been used in such-and-such ways, we consider what differences it made to the achievement of P and this guides our assessment of its overall goodness. Another, quite different rendition of the appraisal view is *normative*. According to this rendition, a model can (and perhaps should) be *made* better, other things being equal, by being *made* more performative, i.e., able to make/actually making larger differences to the achievement of P.

In van Basshuysen et al. (2021, 123), we caution that this normative rendition of the appraisal view is highly problematic, since it invites tuning model forecasts to steer people's behaviors in certain directions. Even if the purposes that modelers sought to promote this way were successfully tracking an uncontroversial set of values V, we should think that this practice is nevertheless highly questionable. Models are widely considered to be epistemic instruments: to the extent that they help with the achievement of practical purposes, this should be only as a function of sincere epistemic contributions that they make, but not by meddling with these contributions to effect specific outcomes. We call violations of this constraint *wishful modeling* (ibid.). In a nutshell, wishful modeling happens when non-epistemic values, e.g., concerning the desirability of certain social outcomes, steer the construction and use of models with the explicit aim of manipulating a target system in a specific way. While it is now widely recognized that non-epistemic values can, should, or necessarily do, often play (legitimate) roles at various stages of scientific inquiry (Biddle 2013; Elliott 2017; Elliott and McKaughan 2014; Douglas 2009; Winsberg 2012), using models to specifically steer people's behaviors would contravene even liberal views on acceptable value influences. The concern that the appraisal view may open the door toward misuses of scientific models becomes even more acute when considering potential damage to public trust in model-based science. Narratives that cast models as engines of persuasion and manipulation for intransparent goals could significantly contribute to the ongoing erosion of public trust in science (cf. Kreps and Kriner 2020). What this suggests, and what we emphasized in van Basshuysen et al. (2021), is that we must shut the door firmly on the normative rendition of the appraisal view. Winsberg and Harvard, however, worry that doing so will be difficult in practice and use an analogy to draw out the problem:

> Imagine holding an annual race in which we tell runners that the goal is to complete 10 km in the fastest possible time, but where, year after year, we award the medals to runners who most quickly reach the 5 km mark. Hopefully it is clear that we cannot neatly separate how runners will be evaluated from what they will eventually adopt as their goal. The same will be true of modelling. If those who judge the suitability, adequacy, or usefulness of a model give it high marks when it succeeds performatively (according to the values of the judges in question), they will be sending the signal that modellers should adopt this goal. (Winsberg and Harvard 2022, 4)

So, even if we managed to ensure that models got their counterfactuals right, and pressed modelers to demonstrate that this is so, counting (some) performative effects as good-making features of models could induce incentives that divert modelers' attention away from epistemic goals such as forecast accuracy toward performative goals, such as issuing forecasts that are likely to steer behaviors in putatively beneficial ways. Modelers' goals, we might insist, should primarily focus on doing as good of an epistemic job as possible, which includes getting counterfactual and actual scenario forecasts right, and the best way to ensure this is to exclude performative effects from consideration in model appraisal. According to Winsberg and Harvard, then, the best way of shutting the door on the normative rendition of the appraisal view is to reject the appraisal view altogether.

So if appraisal is no good, how should we deal with model performativity instead? Let me consider a second strategy for dealing with performativity, the *mitigation view*, which, at face value, promises to evade these concerns, but ultimately falls prey to similar worries.

3.2 The Mitigation View

The mitigation view aims to deal with model performativity by getting rid of it. Starting in the 1950s, economists and political scientists began investigating how publicizing election polling results could alter election outcomes (Grunberg and Modigliani 1954; Simon 1954). Two widely-discussed ways in which this could happen are the *bandwagon* and *underdog effects*.[12] The former captures a case where a candidate A is predicted to win an election against B and, because of this public prediction, more voters than otherwise turn out in support of A. The underdog effect describes the converse: if A is predicted to be ahead in the race, some voters who would have otherwise voted for A end up voting for B, because they want to vote for whoever is behind in the race. In either case, the first-

[12] Though see Guala (2007), who casts "genuine" performativity in narrower outlines that would not recognize these cases.

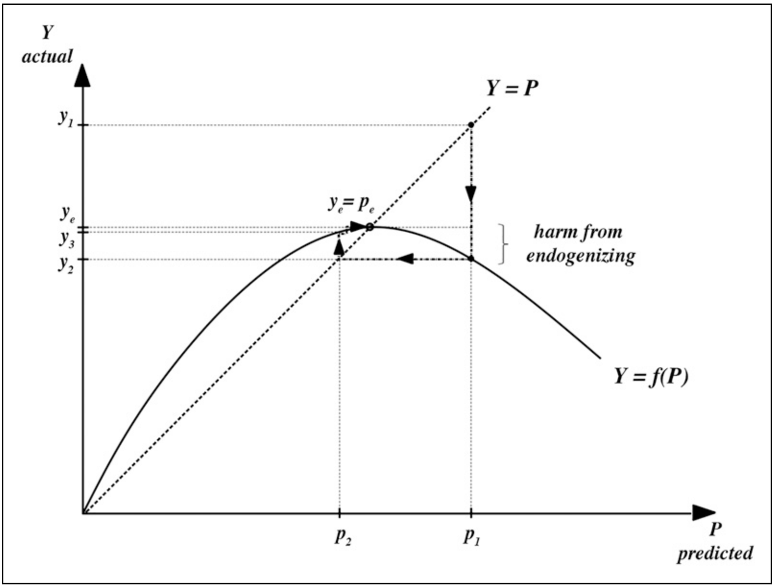

Figure 2. The harm from endogenizing.

stage prediction of the election result will turn out to be inaccurate because agents respond to the public prediction—it is performative.

Against the background of such cases, Grunberg and Modigliani as well as Simon undertook analytical investigations to determine conditions under which public predictions could be modified so as to *endogenize* peoples' behavioral response, i.e., to explicitly model how individuals respond to a prediction and to formulate an adjusted prediction that takes that response into account and brings predictions in line with actual results in equilibrium. Similar efforts to endogenize peoples' behavioral response have recently been undertaken by epidemiological modelers and computer scientists (see Avery et al. 2020b for an overview; see Perdomo et al. 2021 for efforts in machine learning). Figure 2 captures how following the mitigation strategy could look like in the Covid modeling case.

The response curve is the same as before. In addition, the dash-dotted lines and arrows capture schematically how behavioral response is endogenized.[13] The first-stage prediction, P_1, is plugged into a function $f(P)$ that captures the performative response curve. This yields Y_2, which then figures as the second-stage prediction P_2. When publicly predicting P_2, by taking into account how

[13] For simplicity, I trace this out as an iterative process, noting that analytical models may be able to solve for an equilibrium without intermediate stages.

people would have responded to P_1, we move to a different segment of the response curve, which gives us Y_3. In an iterative process of taking the stage-n outcome Y_n to yield the stage-n+1 prediction P_{n+1}, we eventually reach an equilibrium point $Y_e = P_e$ that intersects the 45° line where predictions perfectly coincide with actual outcomes.

Is this a good way of dealing with performativity because it avoids moral and political value judgments encroaching on the appraisal and thereby the construction of scientific models? Not necessarily. Importantly, when taking the mitigation route, Y_e is now higher than Y_2, which would have been the outcome under the appraisal route. So, if we continue to assume for simplicity that minimizing infection numbers is good, endogenizing people's behavioral response is a worse strategy in terms of our practical purposes, but is epistemically superior, since predictions now coincide with actual outcomes.

What is important to note, then, is that there can be *trade-offs* between epistemic and practical purposes and the mitigation view settles this trade-off in favor of epistemic purposes at the potential cost of inferior social outcomes. The appraisal view, by contrast, is open to accepting compromises in predictive performance in exchange for practical benefits. Crucially, this means that *both* routes reflect a value-laden stance on the trade-off, and despite initial appearances to the contrary, the mitigation strategy is subject to concerns about illegitimate value influences, too (see also Brown 2017). Why, after all, should we think that it is overall better to have a model accurately predict infection numbers when this means that those numbers would be higher than if we had not endogenized behavioral response? Why should not we think it can be preferable to have a model overestimating infection numbers (i.e., making first-stage predictions that do not consider performative effects), thereby contributing to behaviors that realize lower numbers? Answering these questions, necessarily, involves moral values because the choice of whether to endogenize or not is not only a choice between better and worse predictive performance but also a choice between two different social outcomes (Y_2 and Y_e). Even if, say, lowering infection numbers were an uncontroversial moral good, and modelers decided to refrain from endogenizing to help achieve this good, it is not obvious that they should: modelers are not suitably legitimized to make choices between social outcomes on our behalf, even if their values magically coincided with a hypothetical aggregate public value profile.

The mitigation view hence leaves us an uncomfortable epistemic-ethical bind. Model performativity can sometimes yield beneficial outcomes, and there are reasons to think that these may be counted toward the overall goodness of a model in epistemic-practical terms. Such a view, however, must also manage concerns about illegitimate value influences that threaten to undermine the epistemic integrity of specific models, and model-based science more generally. An alternative can be to "endogenize away" performative effects by modeling how agents respond to predictions. However, this route similarly faces difficult questions about what legitimizes modelers to make modeling choices that ultimately select different social outcomes than would have been realized without mitigation attempts. Recognizing this helps us appreciate that Winsberg and Harvard's call to reject the appraisal view for the threats posed by its normative rendition does not take us very far when the relevant alternative, mitigation, is subject to similar concerns about illegitimate value influences. Worries that the appraisal view opens the door to such influences are hence misplaced—the door has been open all along, but proponents of mitigation-type approaches have so far not adequately recognized it. Faced with a choice between a rock and a hard place, let me now turn to explore whether we can find some smoother pebbles in between, by considering (1) what principles could help keep the most severe value-related concerns affecting both views at bay and (2) how contextual factors bear on the adequacy of both strategies for managing performativity.

4. Managing Performativity

It is clear now that both routes to deal with model performativity need a safeguarding principle to mitigate concerns about illegitimate value influences at central stages of model construction and use. The main target of such a principle is to prevent moral and political values from illegitimately encroaching on the construction and use of models, be that through modelers tuning models to effect specific behavioral responses or by modelers endogenizing behavioral response and thereby, at least tacitly, making value-related commitments that favor some social outcomes over others. A general principle that could help break these tensions is:

> (***Orthogonality***): modelers' choices with regard to model construction and use should be independent of whether they consider certain performative effects to be desirable.

Addressing the appraisal view, orthogonality requires that the choices modelers make in constructing and using models be robust over changes in their views on the desirability of certain social outcomes. Taking the Covid case, regardless of whether modelers think that saving lives is of utmost importance or that aggressive suppression policies are inappropriate because they unacceptably infringe on civil liberties, their choices of what to model, how to model it, and so on should not be steered by these considerations. Importantly, orthogonality goes both ways and does not only focus on putatively positive performative effects: if a model foreseeably has negative performative effects, this should not influence modeling choices either. For instance, consider a case where policy makers' and the public's priors on the severity of the pandemic early on are wildly exaggerated, anticipating

Millions of deaths within weeks (Y_0), leading to severe lockdowns and individuals choosing to fully self-isolate. Consider now that Covid models are introduced, which forecast much lower numbers (Y_1). In such a case, modelers might expect that policy makers and the public responding to these forecasts could have performative effects that lead to a significant increase in deaths over the aggressive suppression scenario ($Y_2 > Y_1$). This may be considered, by some, a bad-making feature of using the model to inform policy and the public, but orthogonality would insist that this should not influence modelers' choices with regard to how models are constructed, what scenario forecasts are issued, how results are communicated to policymakers and so on (though it might influence decision-makers choices—more on this shortly).

Addressing the mitigation view, orthogonality insists that, in the first instance, the choice of whether to endogenize or not should not depend on modelers' assessment of the desirability of the performative effects to be endogenized. However, orthogonality does not go further than that. It cannot help with the fact that the choice of whether to endogenize or not may invariably promote some social outcomes over others and is therefore, necessarily, a value-laden choice. I expand shortly on how orthogonality may help at least in ensuring decisions such as these are made by agents who are (more) suitably legitimized than modelers, e.g., democratically elected decision-makers.

Whether orthogonality is important, and whether mitigation or appraisal seem more appropriate also depends importantly on contextual factors. Let us consider some cases, including some more fully located in the social science realm, to see this. First, suppose an extreme case where climate modelers across different modeling groups begin to systematically meddle with climate model parameters to exaggerate the extent and projected impacts of anthropogenic climate change. Their aim, let us suppose, is to motivate policy makers and the public to take more drastic action in reducing greenhouse gas emissions and investing in climate change adaptation and mitigation policies. This practice would be highly problematic, for it may undermine the credibility of specific simulation results and of modeling/simulation studies in climate science more generally. Even if the first-order effects on policy and the public's behaviors were desirable as such, promoting performativity in this way should be resisted. Contrast this with a second case involving central banks. Suppose central banks begin to parameterize their macroeconomic forecasting models to induce specific inflation expectations on the part of institutional and private actors so that these behave in ways that help the banks achieve their inflation targets. While certainly questionable, this case does not seem as problematic as the climate science case. First, central banks have a mandate (at least derivatively) to steer inflation expectations, and are widely understood to pursue actions within the envelope of this mandate. Their role is hence significantly different from that of climate scientists, who do not have a mandate to intervene in socioeconomic systems with the aim of effecting, say, specific policy trajectories or certain levels of global warming. A second important difference is that there is a much more level epistemic playing field. Institutional actors (e.g., banks and insurance companies) have their own epistemic resources that can help them resist unduly manipulative attempts on the part of central banks to steer expectations. So, while promoting performative effects seems less problematic in the central bank case for the shared understanding of what central banks' roles are, it will also be substantially less effective, since other actors can more easily tell whether central banks are making epistemically faithful forecasts or are rather trying to steer expectations. What these contextual differences suggest is that performativity is especially problematic when epistemic resources are unevenly distributed, when communicating modeling results to policy makers and the public must rely on a substantial but ultimately fragile architecture of trust, and when agents' shared understanding of scientists' role maintains that they do not, and should not, attempt to steer policy and behavioral response in specific ways. Finally, third, an altogether different case obtains when models themselves, as epistemic tools made available to a population, can have performative effects in shaping outcomes of a target. An interesting case study in this regard are financial market models such as the capital asset pricing models (CAPM) recently discussed by Vergara-Fernández, Heilmann, and Szymanowska (2023). Such models may afford (certain) financial market actors with novel abilities to intervene in the market, including in socially undesirable way, e.g., by creating novel products, institutions, or ways of engaging the market that involve problematic forms of risk imposition on the general public (e.g., increasing the risk of financial and economic crises). Models are epistemic technologies, and when they enable (perhaps foreseeably) negative outcomes by virtue of their performative capacity to re-shape market interactions, we may consider whether modelers who furnish these technologies have special epistemic-ethical responsibilities regarding such effects (Vergara-Fernández, Heilmann, and Szymanowska 2023, 20). Here, orthogonality would seem misplaced, as we may think that value judgments relating to the potential harms induced by releasing models into the wild *should* play an important role in regulating decision-making at various points. Perhaps these should not be modelers' *own* values, but values, as such, are clearly needed and orthogonality should steer clear of ruling out their appropriate involvement.

Despite some progress made on understanding when orthogonality seems important to pursue, it nevertheless remains an extremely general principle, more akin to a goal rather than a recipe for how to achieve that goal. This is a feature shared with many existing attempts in the values-in-science literature to formulate general principles to regulate the ways in which values may bear on the conduct and outputs of scientific research, e.g., versions of the value-free ideal (Betz 2013; Biddle 2013; Brown 2017; Douglas 2009; Elliott 2017). Like such contributions, orthogonality marks an attempt to formulate a general principle which can subsequently be assessed for its plausibility over a range of contexts and refined accordingly. In fine-graining orthogonality, it seems important to consider with regard to what aspects of model use it seems an appropriate goal to pursue, including:

1) How should models be built? What is modeled and how?
2) How should model outputs be interpreted and communicated?
3) What recommendations should be made based on model outputs?
4) How should modeling study results be used in decision-making?
5) How should decisions be communicated to the public?

Across these aspects, orthogonality is only plausible for some and not others. Specifically, it seems that concerns about illegitimate value influences in regards to appraising, facilitating, mitigating, or managing performativity predominantly affect the first two aspects. This is where concerns about wishful modeling are most likely to occur, and where orthogonality is needed.

Turning to stage 3), the issue of whether and how researchers should make recommendations based on model outputs is contentious (Carrier 2022; John 2018; Gundersen 2020). I will not engage here with the various positions that have been offered in the literature and only note that much depends on specifics of the context and what type of view one takes on appropriate roles for scientists advising policy. For instance, when research is specifically commissioned to address concrete evidentiary needs arising in a decision-making context and decision-makers issue a mandate to researchers that allows or instructs them to let specific, independently determined value judgments inform the recommendations they make, this can make orthogonality seem less pertinent. To the extent that decision-makers are suitably legitimized in letting specific values bear on decision-making and researchers are understood to operate as advisors who take pre-determined value-judgments on board when issuing recommendations, orthogonality seems misplaced. Of course, we may still insist that researchers' *own* values should remain orthogonal to what recommendations are made and that they should properly enable the value-influences sanctioned by decision-makers even if these contravene their own values. So, depending on such contextual factors, orthogonality may seem important to pursue, or rather misplaced, and no *general* judgment concerning its appropriateness regarding (3) seems plausible.

Focusing on (4) and (5), it is clear that the decisions at issue here are rarely made by modelers, but usually (or so I assume) by suitably legitimized decision-makers afforded with a mandate to make value-laden decisions on behalf of others. Orthogonality, as articulated earlier, trivially does not apply here, since it was framed in terms of modelers' choices. But we could envision a wider reading of orthogonality, according to which model construction and use should be independent also of users' and decision-makers' views on the desirability of certain performative effects. For instance, there might be concerns about decision-makers specifically commissioning studies that are likely to yield outcomes, which they can use to provoke certain desired responses on the part of the public, e.g., by asking researchers to explicitly focus on modeling worst-case scenarios, or by cherry-picking evidence to help promote certain goals. Yet, aside from playing a role in regulating wishful thinking in regard to what research is commissioned or what studies are considered, it seems that orthogonality is not an especially plausible *general* principle for governing decision-making at (4) and (5). Determining how to use modeling study results in decision-making, making such decisions, and communicating them to the public is fully in the realm of policy-making, so it should neither be surprising that moral and political values play important roles here, nor especially worrying that they do (though the specific values involved will often be contestable). Yet, while the role of orthogonality will be relatively less important in regard to (4) and (5), it should be emphasized that there can nevertheless be more general concerns about performativity that arise here, too.

For one, while we might cynically say that much of public policy making is about steering the public's behaviors in certain, putatively desirable ways, it is important to insist that policy-makers do not misuse models for these purposes. Specifically, decision makers should not claim that the policies they adopt follow straightforwardly from modeling studies, emphasizing that they merely "follow the science." Models may play a justificatory *epistemic* role in helping decision-makers justify beliefs that enter into a decision-making procedure, but they should not play an unmediated justificatory role in recommending specific actions be taken, especially if the values involved in selecting those actions remain poorly articulated. Winsberg and Harvard share this concern when emphasizing that "[…] we should guard against models being used to justify existing political views by representing their favoured policies as the ones that 'follow the science'. Otherwise, our standards of scientific and democratic scrutiny will suffer" (Winsberg and Harvard 2022, 6). Reinforcing these concerns, we might indeed insist on a *dependency* principle, which requires decision-makers to articulate how their decisions depend not only on model forecasts but also on values that are necessarily involved in arriving at decisions that are informed by such forecasts.

A second, related concern is that policymakers acting on a forecast can itself have performative effects by lending further credibility to that forecast (or diminishing it), thus additionally influencing individual's response to forecasts, and it seems unclear what epistemic-ethical obligations policy makers have in regard to referring to specific modeling results as underwriting their decision-making (see Carrier 2022). This relates intimately to broader concerns about deciding whether to communicate model study results at all to the public, which results to communicate, and how to do so. Policy-makers could conceivably choose to withhold forecasts from the public if they believe that publicizing them would have undesirable performative effects. This, of course, can be highly controversial, just like cherry-picking evidence or wishful modeling. Importantly, however, these are controversies about the justification and communication of political decision-making rather than about the construction and use of models *by modelers* and so insisting on orthogonality here does not seem to be the right kind of way to deal with these controversial aspects of public policymaking.

How can orthogonality be achieved in those cases where it is appropriate to pursue? I will not provide detailed recipes here, as this issue hinges significantly on contextual factors, including the stakes involved, the purposes for which models are used, how level the epistemic playing field is, what epistemic and practical resources are available, and so on. So, this is best left for future work. That said, it seems likely that existing proposals to manage the influence of moral and political values on the conduct and outcomes of scientific research provide a fruitful menu of options to explore (see also Godman and Marchionni 2022; Nixon et al. 2022 for related proposals), including measures such as (1) considering what kinds of institutional designs can help facilitate a clearer division of labor in regard to what decisions are made by which types of agents, (2) making progress in further articulating the epistemic-ethical duties of modelers and policy-makers (see Winsberg and Harvard 2022, 6) and exploring what roles modelers may legitimately refrain from playing (e.g., making decisions about how to handle performativity), (3) facilitating transparency and public understanding of the invariably value-laden nature of using models for policy, and (4) promoting open science measures, such as open data, open code, open peer review, pre-registration and pre-analysis plans, and related instruments that prompt researchers to articulate the epistemic and practical purposes of their modeling studies to help promote public scrutiny.

5. Conclusions

Model performativity is a thorny phenomenon that raises important value-related concerns about using models for informing policy and the public. When models have performative capacities, such concerns are unavoidable. In the words of Grunberg and Modigliani, "[…] whenever the agent reacts to the public prediction, the forecaster becomes—however unintentionally—a *manipulator*, since his pronouncement affects the operations performed by the agent upon some variables" (1954, 471; italics in the original). Neither the appraisal view, by insisting on an evaluative interpretation and eschewing a normative one, nor the mitigation view, by promising to endogenize the problem away, can keep such concerns at bay. Both approaches can be appropriate ways to respond to model performativity in some contexts, but both must also be accompanied by strong safeguarding principles and good institutional designs to minimize illegitimate value-influences on central aspects of model construction and use. I have proposed orthogonality as a general constraint on mitigation and appraisal, which aims to ensure that central stages of model construction and use proceed in a way that is independent of modelers' valuations of certain performative effects. Since orthogonality is a goal, rather than a practical recipe, additional work is needed to articulate concrete principles for governing the construction and use of models to inform policy and the public. As my arguments suggest, it can be fruitful to explore, through further case studies, the conditions under which model performativity raises important epistemic-ethical problems, to consider how contextual features bear on whether mitigation or appraisal seem more appropriate, and to explore what institutional designs may help ensure that orthogonality is attained (Nixon et al. 2022). Further pursuing this project, I hope, may support modelers and policy makers in constructing and using models in an ethically and epistemically more responsible fashion.

Acknowledgements

I wish to thank Julian Reiss and Philippe van Basshuysen for their very helpful comments and criticisms, the attendants at the LUH-LMU-Wuppertal Workshop 2022, the audiences at ENPOSS 2022, SPSP 2022, the 12th Braga Meetings on Ethics and Political Philosophy, the ISEC Colloquium at the University of Witten/Herdecke, and at INEM 2021, as well as two anonymous reviewers for their valuable comments and suggestions, which helped significantly improve the ideas developed in this manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

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

References

Avery, Christopher, William Bossert, Adam Clark, Glenn Ellison, and Sara Fisher Ellison. 2020a. "Policy Implications of the Spread of Coronavirus." *National Bureau of Economic Research*, Working Paper 27007. doi:10.3386/w27007

Avery, Christopher, William Bossert, Adam Clark, Glenn Ellison, and Sara Fisher Ellison 2020b. "An Economist's Guide to Epidemiology Models of Infectious Disease." *Journal of Economic Perspectives* 34 (4): 79-104. doi:10.1257/jep.34.4.79

Betz, Gregor. 2013. "In Defence of the Value Free Ideal." *European Journal for Philosophy of Science* 3 (2): 207-20. doi:10.1007/s13194-012-0062-x

Biddle, Justin. 2013. "State of the Field: Transient Underdetermination and Values in Science." *Studies in History and Philosophy of Science* 44 (1): 124-33. doi:10.1016/j.shpsa.2012.09.003

Boldyrev, Ivan, and Alexey Ushakov. 2016. "Adjusting the Model to Adjust the World: Constructive Mechanisms in Postwar General Equilibrium Theory." *Journal of Economic Methodology* 23 (1): 38-56. doi:10.1080/1350178X.2014.1003581

Brown, Matthew J. 2017. "Values in Science: Against Epistemic Priority." In *Current Controversies in Values in Science*, edited by Kevin C. Elliott, and Daniel Steel, 64-78. London: Routledge. doi:10.4324/9781315639420

Buck, Roger. 1963. "Reflexive Predictions." *Philosophy of Science* 30 (4): 359-69. doi:10.1086/287955

Callon, Michel, and Alvin E. Roth. 2021. "The Design and Performation of Markets: A Discussion." *AMS Review* 11: 219-39. doi:10.1007/s13162-021-00216-w

Carrier, Martin. 2022. "What Does Good Science-Based Advice to Politics Look Like?" *Journal for General Philosophy of Science* 53 (1): 5-21. doi:10.1007/s10838-021-09574-2

Cinelli, Matteo, Gianmarco De Francisci Morales, Alessandro Galeazzi, and Michele Starnini. 2021. "The Echo Chamber Effect on Social Media." *PNAS* 119 (9): e2023301118. doi:10.1073/pnas.2023301118

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

Elliott, Kevin C. 2017. *A Tapestry of Values: An Introduction to Values in Science*. Oxford: Oxford University Press.

Elliott, Kevin C., and Daniel J. McKaughan. 2014. "Nonepistemic Values and the Multiple Goals of Science." *Philosophy of Science* 81 (1): 1-21. doi:10.1086/674345

Ferguson, Neil M., Daniel Laydon, Gemma Nedjati-Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, et al. 2020. "Report 9: Impact of Non-Pharmaceutical Interventions (NPIs) to Reduce COVID-19 Mortality and Healthcare Demand." Imperial College COVID Response Team. Published March 16, 2020. https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-9-impact-of-npis-on-covid-19/

Friedman, Joseph, Patrick Liu, Christopher E. Troeger, Austin Carter, Robert C. Reiner Jr, Ryan M. Barber, James Collins, et al. 2021. "Predictive Performance of International COVID-19 Mortality Forecasting Models." *Nature Communications* 12: 2609. doi:10.1038/s41467-021-22457-w

Friedson, Andrew I., Drew McNichols, Joseph J. Sabia, and Dhaval Dave. 2020. "Did California's Shelter-in-Place Order Work? Early Coronavirus-Related Public

Health Effects." NBER Working Paper No. 26992. National Bureau of Economic Research, Cambridge, MA. doi:10.3386/w26992

Fuller, Jonathan. 2021. "What are the COVID-19 Models Modeling (Philosophically Speaking)?" *History and Philosophy of the Life Sciences* 43: 47. doi:10.1007/s40656-021-00407-5

Godman, Marion, and Caterina Marchionni. 2022. "What Should Scientists Do about (Harmful) Interactive Effects?" *European Journal of Philosophy of Science* 12: 63. doi:10.1007/s13194-022-00493-7

Guala, Francesco. 2007. "How to Do Things with Experimental Economics." In *Do Economists Make Markets? On the Performativity of Economics*, edited by Donald MacKenzie, Fabian Muniesa, and Lucia Siu, 128-62. Princeton, NJ: Princeton University Press. doi:10.1515/9780691214665-007

Gundersen, Torbjørn. 2020. "Value-Free yet Policy-Relevant? The Normative Views of Climate Scientists and Their Bearing on Philosophy." *Perspectives on Science* 28 (1): 89-118. doi:10.1162/posc_a_00334

Grunberg, Emile, and Franco Modigliani. 1954. "The Predictability of Social Events." *Journal of Political Economy* 62 (6): 465-78. doi:10.1086/257604

Henshel, Richard L. 1993. "Do Self-Fulfilling Prophecies Improve or Degrade Predictive Accuracy? How Sociology and Economics Can Both Be Right." *The Journal of Socio-Economics* 22 (2): 85-104. doi:10.1016/1053-5357(93)90017-F

Ioannidis, John P. A., Sally Cripps, and Martin A. Tanner. 2020. "Forecasting for COVID-19 Has Failed." *International Journal of Forecasting* 38 (2): 423-38. doi:10.1016/j.ijforecast.2020.08.004

Jiménez-Buedo, Maria. 2021. "Reactivity in Social Scientific Experiments: What Is It and How Is It Different (and Worse) than a Placebo Effect?" *European Journal for Philosophy of Science* 11: 42. doi:10.1007/s13194-021-00350-z

John, Stephen. 2018. "Epistemic Trust and the Ethics of Science Communication: Against Transparency, Openness, Sincerity and Honesty." *Social Epistemology* 32 (2): 75-87. doi:10.1080/02691728.2017.1410864

Kreps, Sarah E., and Douglas L. Kriner. 2020. "Model Uncertainty, Political Contestation, and Public Trust in Science: Evidence from the COVID-19 Pandemic." *Science Advances* 6 (43): eabd4563. doi:10.1126/sciadv.abd4563

MacKenzie, Donald. 2006. *An Engine, Not a Camera: How Financial Models Shape Markets*. Cambridge, MA: MIT Press.

Maki, Uskali. 2013. "Performativity: Saving Austin from Mackenzie." In *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*, edited by Vassilios Karakostas, and Dennis Dieks, 443-53. Dordrecht, the Netherlands: Springer. doi:10.1007/978-3-319-01306-0_36

Nixon, Kristen, Sonia Jindal, Felix Parker, Nicholas G. Reich, Kimia Ghobadi, Elizabeth C. Lee, Shaun Truelove, and Lauren Gardner. 2022. "An Evaluation of Prospective COVID-19 Modelling Studies in the USA: From Data to Science Translation." *Lancet Digital Health* 4 (10): E738-47. doi:10.1016/S2589-7500(22)00148-0

Northcott, Robert. 2022. “Reflexivity and Fragility.” *European Journal for Philosophy of Science* 12: 43. doi:10.1007/s13194-022-00474-w

Parker, Wendy S. 2020. “Model Evaluation: An Adequacy-for-Purpose View.” *Philosophy of Science* 87 (3): 457-77. doi:10.1086/708691

Perdomo, Juan. C., Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2021. “Performative Prediction.” arXiv:2002.06673v4.

Schroeder, S. Andrew. 2021. “How to Interpret Covid-19 Predictions: Reassessing the IHME’s Model.” *Philosophy of Medicine* 2 (1): 1-7. doi:10.5195/pom.2021.43

Sears, James, J. Miguel Villas-Boas, Vasco Villas-Boas, and Sofia Berto Villas-Boas. 2023. “Are We #StayingHome to Flatten the Curve?” *American Journal of Health Economics* 9 (1): 71-95.

Simon, Herbert A. 1954. “Bandwagon and Underdog Effects and the Possibility of Election Predictions.” *Public Opinion Quarterly* 18 (3): 245-53. doi:10.1086/266513

Tee, Sim-Hui. 2019. “Constructing Reality with Models.” *Synthese* 196 (11): 4605-22. doi:10.1007/s11229-017-1673-8

Van Basshuysen, Philippe, Lucie White, Donal Khosrowi, and Mathias Frisch. 2021. “Three Ways in which Pandemic Models May Perform a Pandemic.” *Erasmus Journal for Philosophy and Economics* 14 (1): 10-127. doi:10.23941/ejpe.v14i1.582

Van Basshuysen, Philippe. 2022. “Austinian Model Evaluation.” preprint. http://philsci-archive.pitt.edu/id/eprint/21233

Vergara-Fernández, Melissa, Conrad Heilmann, and Marta Szymanowska. 2023. “Contextualist Model Evaluation: Models in Financial Economics and Index Funds.” *European Journal for Philosophy of Science* 13: 6. doi:10.1007/s13194-022-00506-5

Winsberg, Eric. 2012. “Values and Uncertainties in the Predictions of Global Climate Models.” *Kennedy Institute of Ethics Journal* 22 (2): 111-37. doi:10.1353/ken.2012.0008

Winsberg, Eric, Jason Brennan, and Chris W. Surprenant. 2020. “How Government Leaders Violated Their Epistemic Duties during the SARS-CoV-2 Crisis.” *Kennedy Institute of Ethics Journal* 30 (3–4): 215-42. doi:10.1353/ken.2020.0013

Winsberg, Eric, and Stephanie Harvard 2022. “Purposes and Duties in Scientific Modelling.” *Journal of Epidemiology and Community Health* 76: 512-17. doi:10.1136/jech-2021-217666

Author Biography

Donal Khosrowi is a postdoctoral researcher at Leibniz University Hannover, Germany. His research interests and recent publications focus on the epistemology and ethics of artificial intelligence, causal inference in social science (especially in economics and evidence-based policy), scientific representation with models, and values in science issues.