Philosophy of Science (2026), **00**, 1–20
doi:10.1017/psa.2026.10306

PSA PHILOSOPHY OF SCIENCE ASSOCIATION
PHILOSOPHY OF SCIENCE

ARTICLE

The One-Step Principle: Performative Model Selection under Perceived Admissibility Parity

Alessandro Guerra

Department of Philosophy, University of Milan, Milan, Italy
Email: alessandro.guerra@unimi.it

(Received 02 June 2026 UTC; revised 25 September 2026 UTC; accepted 25 September 2026 UTC)

Abstract

Scientific models can be performative: Their uptake may alter the phenomena they represent. Existing debates on performativity have not provided a criterion for model selection when models perceived to be in admissibility parity differ in their foreseeable performative effects. Using game-theoretic models of proenvironmental behavior as a test case, I introduce the One-Step Principle: a conservative constraint on performative model selection. The principle permits replacing a pessimistic model with a more optimistic alternative only under conditions of salience, favorable conditions for self-fulfilling performativity, and proximity to the model it replaces.

1. Introduction

As is increasingly recognized in philosophy of science, especially in debates on the human and social sciences, models can be performative:[1] Once published, taught, or used in policy discourse, they may reshape agents' expectations and thereby influence the very phenomena they are meant to describe (cf. Marchionni et al. 2024; van Basshuysen 2025). Model performativity has therefore often been regarded as a challenge to both the epistemic authority and the ethical integrity of science. Epistemically, performativity makes model evaluation more difficult: When agents react to a model, it may be unclear whether the model has captured a preexisting pattern or helped bring that pattern about. From an ethical point of view, performativity raises questions about modelers' responsibility for the downstream effects of their representations, especially when these effects are foreseeable and socially significant.

[1] I use "performativity" in the sense mainly developed in philosophy of economics and economic sociology, where theories or models may contribute to shaping the phenomena they describe. Marchionni et al. (2024) classify model performativity as one form of the broader phenomenon of reactivity, alongside looping effects, the Hawthorne effect, observer effects, experimenter effects, and experimenter demand effects.

© The Author(s), 2026. Published by Cambridge University Press on behalf of Philosophy of Science Association. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

Existing responses to this problem pull in different directions. One cautious response is to deny that the desirability of performative effects should directly guide model choice. On this view, modelers should clarify the purpose of their models, assess whether they are adequate for that purpose, and avoid disguising attempts to shape behavior as neutral representation (Winsberg and Harvard 2022). Related strategies recommend incorporating agents' responses into models where possible, assessing the effects of performative models after their uptake, or keeping the desirability of such effects orthogonal to modeling choices (van Basshuysen et al. 2021; Khosrowi 2023). A more permissive line, by contrast, holds that performativity need not always be treated merely as a threat. Because model uptake may already affect the social world, refusing to consider foreseeable performative effects is not obviously neutral (Guala 2007; Godman and Marchionni 2022). Ortmann's recent defence of performative intervention develops this permissive line in a stronger form, arguing that deliberately shaping target systems through scientific representations need not always be paternalistic or illegitimate (Ortmann 2025).

However, the existing debate still lacks a sufficiently precise criterion for deciding when one admissible model may be replaced by another because of their different foreseeable performative effects. In particular, what remains underdeveloped is a constrained account of permissible model replacement under perceived admissibility parity. This gap matters because, without such a criterion, the debate risks oscillating between two unsatisfactory options: excluding foreseeable performative effects from model choice altogether, or allowing them to guide selection in an insufficiently constrained way. This article addresses the gap by asking when one admissible representation may permissibly be replaced by another with more favorable performative effects, when the available evidence does not clearly favor either and both remain adequate for the purpose at hand. I answer this question by providing a conservative replacement rule that allows performative considerations to matter without making them independently sufficient for model choice.

I develop this account through the case of game-theoretic models of proenvironmental behaviors (PEBs). This case is especially useful because environmental behavior can often be represented by more than one idealized strategic model, while the public uptake of such models may affect the expectations on which cooperation depends. In many PEB contexts, the available evidence does not straightforwardly determine whether noncooperation is driven by a dominant incentive to defect, by lack of assurance about others' cooperation, or by uncertainty about whether others will coordinate on a shared norm. The same behavioral pattern may therefore be compatible with different strategic interpretations.

A PEB interaction may be framed as a Prisoner's Dilemma (PD), where defection appears individually dominant; as a Stag Hunt (SH), where cooperation is possible but depends on mutual assurance; or as a Hi-Lo (HL), where agents' interests are largely aligned, and the problem is mainly one of coordination. Because these representations make different expectations and obstacles to cooperation salient, their public uptake may have different performative effects. In this setting, the general question becomes when a more pessimistic model, such as PD, may permissibly be replaced by a more optimistic alternative, such as SH or HL (Kopec 2017; Ortmann and Veit 2023).

To guide model choice in such cases, I propose the One-Step Principle. The principle regulates a specific form of model selection: the replacement of a

currently pessimistic model, such as PD, with a more optimistic alternative, such as SH or HL. Replacement is permissible only under three conditions—perceived admissibility parity, model salience, and favorable conditions for self-fulfilling performativity—and one procedural restriction: It must proceed by one step, from the currently salient model to the nearest admissible model in the normatively preferred direction. The question of who has standing to fix that direction—and how the risk of manipulation is contained—is addressed in section 4.4.

The principle occupies an intermediate position between the responses developed to deal with performative cases. Against the more cautious view, I argue that foreseeable performative effects need not always be excluded from model selection. In some cases, they may legitimately count in favor of replacing one admissible model with another. But, against an unconstrained consequentialist use of performativity, I also argue that modelers and policy makers may not simply select the model whose uptake is expected to produce the most desirable outcome. Performative effects may permissibly inform model replacement only when the principle's conditions are satisfied.

The article proceeds as follows. Section 2 introduces the philosophical debate on model performativity. Section 3 analyzes the performativity of game-theoretic models, focusing on PD, SH, and HL. Section 4 specifies the scope of the argument, addresses the legitimacy of the model user and the risk of manipulation, and formalizes the One-Step Principle as a criterion for permissible replacement under perceived admissibility parity. Section 5 discusses the principle as a conservative strategy for performative model selection. Section 6 concludes.

2. Models' performativity

The notion of performativity has been discussed across philosophy of science, economics, and the social sciences (Buck 1963; Henshel 1993). Building on these traditions, Khosrowi (2023) proposes a general characterization according to which 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" (374). According to this definition, performativity is target-relative: A model is not performative merely because it has causal effects, but because its use or uptake can affect the very phenomenon it is meant to represent. Whether this capacity is realized depends on causal pathways connecting the model, its users, and the relevant target system.

Model performativity can take many forms, but the present discussion focuses on cases in which a model's use or communication changes the expectations and behaviors of agents in the target system. In some cases, this occurs through explicit predictions; in others, through the representation of a situation as having a certain strategic or social structure. Among the different forms of performativity, self-fulfilling and self-effacing performativity are particularly significant. Self-fulfilling performativity obtains when a model's uptake helps bring about, or stabilize, the state of affairs it represents. Self-effacing performativity obtains when the model's uptake prevents or undermines the represented situation (Khosrowi 2023, 374–75). This distinction is important because performative models can affect their targets in different directions: They may make a represented pattern more likely, but they may also contribute to its disappearance.

Merton's classic discussion of self-fulfilling prophecies illustrates the general mechanism at stake. In his bank-run example, a false belief that a solvent bank is failing leads depositors to withdraw their money, thereby helping to bring about the collapse they feared (Merton 1948). In scientific and economic modeling, an analogous mechanism arises when the uptake of a model changes the target system it is used to represent. The Black–Scholes formula for option pricing provides a paradigmatic example: MacKenzie (2006) argues that its repeated use by financial agents contributed to making option prices conform more closely to the model's predictions. The model's empirical success was, at least in part, the result of its own performative effects.

Guala (2007) analyzes performativity in experimental economics and distinguishes between two mechanisms through which economic models can shape their targets: by informing the design of institutions and by directly shaping individual behavior within those institutions. Although the debate on performativity has historically been centered on economics, the phenomenon is not domain-specific. Van Basshuysen and colleagues (2021), for instance, argue that pandemic models influenced the course of COVID-19 by shaping scientific advice, public policy, and individual behavior.

Taken together, these cases show that model performativity can alter empirical fit and generate normatively significant consequences. Whether these consequences are epistemically or ethically problematic depends on the context of application and on the values at stake. But once performative effects are foreseeable, they cannot be treated as merely accidental by-products of modeling practice. They raise the further normative question of when, and under which constraints, they may permissibly inform model selection.

3. The performativity of environmental behavior models

The environmental case provides a structured test case for the problem of performative model selection. Climate cooperation has often been represented as a Prisoner's Dilemma (PD), a game-theoretic model in which defection is individually dominant and mutual noncooperation is the unique Nash equilibrium[2] (Kopec 2017; Ortmann and Veit 2023). Yet some environmentally relevant interactions may also be represented, depending on agents' incentives and expectations, as Stag Hunt (SH) or Hi-Lo (HL) games, where cooperation is not ruled out by the strategic structure (Asker and Stefánsson 2023; Ortmann and Veit 2023).

These models are useful for present purposes because they form a simple ordered sequence from conflict to coordination and because their public use may shape how agents understand the obstacles to cooperation. The relevant case is one in which modelers or policy makers use or communicate a game-theoretic representation of an environmental interaction. Depending on the model selected, agents may come to see noncooperation as individually rational, cooperation as dependent on mutual assurance, or coordination on a shared norm as feasible. Different admissible representations may therefore have different performative effects. This section explains how PD, SH, and HL represent different obstacles to cooperation and how their uptake may shape the expectations producing performative effects.

[2] A Nash equilibrium is a strategy profile in which no player can improve her payoff by unilaterally changing strategy, given the strategies chosen by the others.

		Player 2	
		C	***D***
Player 1	***C***	(3, 3)	(0, 5)
	D	(5, 0)	(1, 1)

Figure 1. Prisoner's Dilemma.

3.1. Game-theoretic models

One of the most famous models used to represent environmental strategic interactions is the Tragedy of the Commons (ToC), which was first introduced by William Forster Lloyd (1833), and later popularized by Garrett Hardin (1968) in his homonymous paper. In a ToC, individually rational exploitation of a shared resource leads to collectively harmful outcomes when no adequate rules or institutions constrain use. Each agent has an incentive to increase their own benefit from the common resource, while the costs of overuse are distributed across all users. The result is a conflict between individual rationality and collective welfare.

The ToC provides a clear image of the situation among the actors[3] involved in climate change mitigation and this framing is echoed in the IPCC's 2014 assessment (Kopec 2017, 208). On the one hand, each actor has an interest in emitting greenhouse gases as much as possible because the more they can emit, the more they can produce. On the other hand, the effective well-being of an actor (and its effective possibility to gain resources in the future) depends on what others do, because if everyone chooses to maximize their emissions, the result is the ruin of the common. The logic behind the ToC is usually considered as that of the PD, the model that best captures the conflict between individual rationality and collective welfare (Figure 1).[4]

The PD involves two players who must independently decide whether to cooperate (C) or defect (D). The structure of payoffs satisfies the inequality $T > R > P > S$, where T is the temptation payoff (defecting while the other cooperates), R the reward for mutual cooperation, P the punishment for mutual defection, and S the sucker's payoff (cooperating while the other defects). A common representation is shown as follows.

Here, $T = 5$, $R = 3$, $P = 1$, $S = 0$. Given this structure, defection is the dominant strategy for both players: Each achieves a better outcome by defecting, no matter what the other does. DD is the only Nash equilibrium of the game. Yet, when both defect, the collective result $(1 + 1)$ is worse than if both had cooperated $(3 + 3)$. This paradox captures how individually rational decisions can produce socially inefficient outcomes.

For example, suppose that an individual is aware of the public goal established by their country to reduce its emissions, and is therefore uncertain about whether to take a short-haul flight to spend a weekend in London. The uncertainty arises from the fact that a single action of this kind is unlikely to have a significant impact on the

[3] The generic term "actor" allows us to consider climate change mitigation at different levels: individuals, organizations, countries.

[4] There is no consensus in the literature on the game-theoretic representation of the ToC. For an overview of this issue and a compelling explanation of how PD represents the ToC, see Carrozzo Magli et al. (2021).

		Player 2	
		C	***D***
Player 1	***C***	(4, 4)	(0, 3)
	D	(3, 0)	(2, 2)

Figure 2. Stag Hunt.

country's emissions. Yet, if such behavior were to become widespread, the cumulative effect would be highly detrimental to the achievement of the goal. As in the PD, defection (taking the flight) dominates cooperation (not taking the flight) from an individual standpoint, even though the cooperation of all citizens would contribute significantly to achieving the goal.[5]

However, not all environmentally relevant interactions exhibit this kind of incentive conflict. In many cases, the central difficulty is not the temptation to free-ride, but rather the problem of coordinating on a mutually beneficial outcome under conditions of strategic uncertainty. These situations are more adequately represented by coordination games. The Stag Hunt (SH) is the paradigmatic game-theoretic model of this second class of interactions. In contrast to the PD, SH does not involve a dominant strategy to defect: Cooperation is already a Nash equilibrium, but it is fragile because its stability depends on agents' expectations about one another's behavior. The strategic problem is therefore not to override incentive conflict, but to secure mutual assurance. The SH can be represented by the following payoff matrix (Figure 2).

In this case, $R = 4$, $T = 3$, $P = 2$, $S = 0$. Two Nash equilibria emerge: one cooperative (CC) and one risk-dominant but inefficient (DD). Mutual cooperation provides the highest collective payoff ($4 + 4$) but is only stable if both players expect the other to cooperate. When expectations are uncertain, the safer equilibrium (DD) may prevail.

A paradigmatic instance of a PEB that can be modeled as a SH game is the decision to install solar panels within a residential community. The situation captures the essential features of coordination problems where collective success depends on mutual participation. Individuals face a choice between cooperating (installing solar panels) and defecting (refraining from installation). The benefits of clean energy arise only if a critical mass of individuals cooperate. Conversely, isolated cooperation yields little benefit and entails higher private costs.

Although, to the best of my knowledge, it has not appeared in the literature on PEBs, situations in which individual tension is low can be effectively represented using the Hi-Lo game. The HL game represents a cooperative form of social interaction in which agents share aligned interests. In contrast to the PD and SH, HL involves no conflict of interest. All players prefer the same outcome, and coordination failure arises solely from uncertainty about others' behavior. The game can be represented with the following matrix (Figure 3).

The outcome CC is usually achieved because each player holds the same preference for a higher payoff, and there is no temptation to defect. However, if one player has

[5] Asker and Stefánsson (2023, 270) present a similar table illustrating individuals' decisions to take a Sunday drive in a gas-guzzling SUV.

		Player 2	
		C	***D***
Player 1	***C***	(4, 4)	(0, 0)
	D	(0, 0)	(2, 2)

Figure 3. Hi-Lo.

reasons to believe that the other will defect, it is rational to defect. The game has therefore two Nash equilibria, namely CC and DD.

For instance, refraining from littering in public spaces can be modeled as a HL. This behavior is typically rooted in norms that make it common—and considered right—not to litter in shared environments. Everyone benefits from clean shared environments, and the personal gain from littering is typically limited. Nevertheless, if individuals doubt that others will comply, the norm can erode, and littering may spread. The challenge is thus to maintain mutual expectations that sustain the cooperative equilibrium.

Taken together, the PD, SH, and HL games delineate a conceptual continuum from conflict to cooperation in individuals' engagement in PEBs. This continuum should not be understood as mechanically fixed by the type of behavior alone. Different agents may evaluate the costs and benefits of the same PEB differently, and the same environmental behavior may therefore be more plausibly represented by different games in different contexts.

For this reason, the role of PD, SH, and HL in this article is not to provide a complete representation of environmental behavior.[6] They function as idealized templates that make the problem of performative model selection especially tractable. Their simplicity is an advantage for present purposes: They differ in a clear and ordered way with respect to the obstacle they identify to cooperation—dominant incentives to defect, lack of mutual assurance, or mere coordination failure. They therefore serve as a test case for formulating the One-Step Principle developed in section 4. More specifically, the case considered is one in which modelers or policy makers must decide which admissible representation to communicate, while having reason to expect that different representations may shape agents' expectations differently.

3.2. Game-theoretic performativity

Game-theoretic models can have at least three distinct functions in the analysis of PEBs. First, they are descriptive: They represent the strategic structure of interactions among agents facing environmental choices. Second, they are normative: Given a certain distribution of payoffs, they identify which actions are instrumentally rational in light of agents' beliefs about one another's behavior. Third, they can be

[6] The use of one-shot games here is an idealizing device. Many environmentally relevant interactions are repeated, and repetition may affect the credibility of cooperative expectations. In repeated PD, cooperation may become sustainable under familiar Folk Theorem conditions when agents sufficiently value future interaction and when reputational, monitoring, or sanctioning mechanisms are available (Fudenberg and Maskin 1986; Mailath and Samuelson 2006). Repeated-game models are not treated here as a separate class of candidates.

performative: When disseminated and taken up by agents, they may shape beliefs and expectations in ways that causally affect the very interactions they aim to describe.

In the game-theoretic setting, model performativity operates primarily through expectations. A model is performative when its use and communication help stabilize or undermine particular patterns of behavior in its target system. Game-theoretic models are particularly well suited to making this mechanism explicit because they already treat agents' expectations about others' actions as a determinant of equilibrium selection.

To make this link precise, it is useful to draw on the Lewis (1969)–Schelling (1960) account of conventions as equilibria of coordination games. On this view, a coordination game is a strategic situation in which agents benefit from coordinating their actions, prefer coordinated outcomes to uncoordinated ones, and must form reciprocal expectations to select among multiple possible equilibria.

Building on the Lewis–Schelling framework, Guala (2016) connects conventions and performativity: Conventions are stable patterns of behavior sustained by shared expectations, and changes in expectations can therefore alter the convention. Because coordination is often guided by focal points or coordination devices, publicly communicated models can become performative when they make a particular interpretation of the strategic situation salient. They do so by shaping how agents conceptualize the interaction and which equilibria they regard as focal.

This perspective is particularly useful for understanding the performativity of SH and HL models of PEBs. Both are coordination games with multiple Nash equilibria, so model uptake can affect which equilibrium agents regard as focal. In SH, this problem is especially acute: Cooperation is an equilibrium, but it is fragile because agents have reason to defect when they doubt others' willingness to cooperate. In HL, by contrast, interests are aligned, and the higher-payoff equilibrium is comparatively focal.

PD raises a different problem. In the standard one-shot analysis, PD is a mixed-motive game with a unique Nash equilibrium in mutual defection. Because there is no cooperative equilibrium to be selected, the Lewis–Schelling mechanism cannot explain PD performativity in the same way as it explains SH or HL performativity. A PD representation may nevertheless be performative in a different sense: By making defection appear strategically unavoidable, it may stabilize pessimistic expectations and discourage the development of the mechanisms—such as repeated interaction, reputation, monitoring, credible commitments, or institutional support—that could make cooperation sustainable (Fudenberg and Maskin 1986; Mailath and Samuelson 2006).

The upshot is that PD, SH, and HL differ not only in their formal structure but also in their performative profiles. PD may foreground defection as the rational response to conflict of interest; SH may present mutual assurance as the central problem; and HL may frame coordination on an already aligned outcome as the relevant task. This ordered contrast provides the representational basis for the One-Step Principle developed in the next section.

In what follows, I focus on self-fulfilling rather than self-effacing performativity. The relevant question is whether the uptake of a model may help stabilize the expectations and behaviors that the model makes salient. Self-effacing cases, in which uptake undermines the pattern represented, raise a different problem and fall outside the scope of the present argument.

4. The One-Step Principle

This section develops the One-Step Principle. Section 4.1 situates the principle within existing responses to performative model choice, and explains why a further, more restrictive criterion is still needed. Section 4.2 introduces the notion of perceived admissibility parity. Section 4.3 specifies the scope of the principle: the class of models to which it applies. Section 4.4 addresses who has standing to treat one admissible model as preferred, and how the risk of manipulating the agents affected by that choice is contained. Section 4.5 states the principle formally.

4.1. The normative problem and its existing responses

When two or more models are admissible but likely to differ in their performative effects, model choice becomes ethically significant. Deliberately taking such effects into account may appear to license a problematic form of steering: Modelers or policy makers would no longer merely represent a target system but would use representations to influence agents' behavior. This is the kind of worry raised by Winsberg and Harvard (2022): If models are selected or communicated to produce desirable behavioral effects, scientific representation risks being turned into a disguised form of intervention.[7] Several responses in the literature attempt to avoid this difficulty by limiting, neutralizing, or redescribing the role of performativity in scientific modeling. These responses can be grouped into four strategies.

The first strategy is ex post appraisal. On this view, modelers should not select models in advance because of their expected behavioral consequences, but they may later assess whether those consequences were beneficial. This strategy is associated with van Basshuysen et al.'s (2021) discussion of pandemic models, where models with limited predictive accuracy might still be evaluated positively if their uptake contributed to desirable behavioral responses. The second strategy is orthogonality. Khosrowi's (2023) principle of orthogonality holds that modeling choices should not depend on whether the performative effects of a model are desirable. Orthogonality therefore protects model choice from being directly governed by modelers' preferences. The third strategy is endogenization, discussed by van Basshuysen et al. (2021) and Khosrowi (2023). Rather than using performativity as a practical tool, modelers may try to incorporate agents' responses to the model into the model. This treats performative feedback as an object of explanation rather than as something to be exploited. The fourth strategy is conditionalization. Instead of presenting unconditional forecasts, modelers may offer claims of the form: If agents do X, Y will follow. This appears to avoid paternalism because it preserves agents' choice among possible courses of action. This strategy is central to Winsberg and Harvard's (2022) response to worries about performative modeling.

Ortmann's (2025) critique is that none of these strategies fully removes the normative burden generated by performativity. Ex post appraisal might be hypocritical when performative effects are foreseeable ex ante: If modelers have reason to expect that their models will influence behavior, refusing to consider those effects does not make model choice neutral. Orthogonality rightly warns against

[7] This problem has recently been discussed under the heading of performative paternalism. Ortmann (2025) offers a systematic formulation of this worry.

letting modelers' political preferences determine model choice, but it becomes too restrictive if foreseeable performative effects are ethically relevant. Endogenization may be epistemically valuable, but it may also be practically unavailable when behavioral mechanisms are complex, unstable, or data-poor. Finally, conditional projections do not always stand outside performativity. In social contexts, communicating a conditional projection can alter expectations and behavior; a projection about a bank run or climate cooperation may become part of the causal process it describes.

Ortmann's claim is that performative intervention is not inherently paternalistic or illegitimate.[8] I take this conclusion to be correct, but incomplete for the problem considered here. The issue is not simply whether performative effects may ever be taken into account. It is how they may permissibly guide model replacement once more than one model has already satisfied the relevant admissibility threshold (section 4.2).

The position developed here therefore combines Ortmann's permissive insight with the caution emphasized by Winsberg and Harvard (2022). What is needed is a rule that identifies which replacement, if any, is permissible within an already restricted set of admissible models. The PD, SH, and HL models of PEBs provide a tractable case in which this question becomes especially clear: When a pessimistic but admissible representation and a more optimistic but also admissible alternative differ in their expected performative effects, the normative question is whether the latter may replace the former. The One-Step Principle addresses this question by providing the conservative rule for model replacement formulated in the next subsections and defended in section 5.

4.2. Perceived admissibility parity

The One-Step Principle applies only under a specific condition: perceived admissibility parity. This obtains when two or more models are admissible for a given modeling purpose and the available evidence does not clearly favor one of them. Admissibility is a threshold notion: A model either clears it or it does not. There is parity among models that clear this threshold.

The threshold has two components. The evidential component asks whether the available behavioral or preference evidence is compatible with the structure a model attributes to the target system. The purpose-relative component asks whether the model's resolution is fine enough to distinguish the mechanisms on which the decision at hand turns. This draws on the more general point that a model's adequacy is not intrinsic to the model, but relative to the purpose for which it is used (Parker 2020). The relevant purpose here is representational: capturing, with sufficient accuracy for the decision at hand, the target system agents face. A model clears the threshold if both the components are satisfied: Its structure is not contradicted by the evidence, and its resolution is not coarser than the decision requires.

Whether a model clears this threshold is assessed through a judgment made by a relevant model user. In scientific contexts, this will typically be the modeler; in policy

[8] Some expectation-shaping interventions are overt, authorized, and compatible with the autonomy of those affected. Central banks, for instance, may use public communication to influence inflation expectations because shaping expectations is part of their institutional role (Ortmann 2025).

contexts, it includes those responsible for adopting or communicating the model, namely policy makers and the public bodies that communicate it to the public. The model user is also responsible for establishing the normatively preferred direction—that is, whether a more favorable model, M^+, may replace a currently salient one, M^-, in the sense formalized in section 4.5. This determination, however, is not thereby legitimate: The conditions under which it becomes so are addressed in section 4.4.

Consider, for example, a local campaign aimed at reducing short-haul leisure flights. Suppose that individuals recognize the environmental costs of flying but expect others to continue flying, while also reporting a conditional willingness to abstain if sufficiently many others do the same. The evidence is compatible with agents facing a dominant incentive to defect, as in a PD, and equally compatible with agents being willing to cooperate but lacking assurance about others' cooperation, as in a SH. Both models therefore clear the evidential component, and neither is disqualified by an insufficiently fine resolution: Each identifies a distinct, decision-relevant reason for the observed noncooperation. In such a case, perceived admissibility parity obtains between PD and SH. By contrast, a model that collapsed both patterns into a single undifferentiated category of "noncooperation" would fail the purpose-relative component: However well it fit the evidence, it could not tell the campaign's organisers whether to address free-riding or assurance, which is exactly what the decision at hand requires distinguishing.

A similar indeterminacy arises in climate negotiations. As DeCanio and Fremstad (2013) argue, negotiating records remain compatible with several distinct game-theoretic characterizations because "which game corresponds to the actual state of the world depends both on the severity of risks associated with climate change and the perceptions of the governments engaged in the negotiations" (177). In either case, perceived admissibility parity is ex ante and fallible: It does not imply that admissible alternatives are equally accurate. However, despite the possibility of being in error from the descriptive point of view, the choice in this kind of situation is particularly relevant in virtue of the likely performative effects.

4.3. The scope

The One-Step Principle is intended to cover a specific class of performative models, having the following characteristics. First, the candidate models must be in perceived admissibility parity. Second, the admissible models must form a discrete, ordered family—as opposed to a continuum of representations—along a normatively relevant dimension, so that a relation of proximity between neighbouring models ("a step") is well defined, and models' performativity must operate through this same discrete family: Uptake must move agents to converge on one of the discrete alternatives.

A case in point, detached from the environmental case, is bank runs, already introduced in section 2 by Merton (1948). Suppose the eventual number of withdrawing depositors is fixed, but three admissible models disagree about the mechanism governing the timing of withdrawal. In the first, each depositor withdraws out of fear of being last, regardless of what others do, producing a rapid, unconditional run. In the second, withdrawal is conditional on observing a threshold number of others already withdrawing, producing a run whose pace depends on mutual assurance. In the third, withdrawal timing reflects each depositor's own

liquidity needs, largely independent of others' behavior, producing a slow, uncoordinated pattern. These three accounts are ordered from most to least pessimistic, from the standpoint of how quickly and how certainly the bank becomes insolvent—and moving from the first directly to the third would skip the second, just as a shift from PD to HL would skip SH.

The same evidence available to a model user about depositors' likely behavior may be compatible with more than one of these accounts, and because the three accounts differ in mechanism—not merely in the rate at which withdrawal proceeds—uptake moving depositors to converge on one of them rather than another can affect how quickly insolvency occurs.

By contrast, many performative cases do not involve choosing among a family of alternative models at all. The Imperial College CovidSim, an epidemiological model of the COVID-19 pandemic, is a case in point. Here, uptake is typically performative through continuous feedback on a single model's aggregate parameters, not through a shift from one admissible model to another. A higher projected infection rate makes people somewhat more cautious, but this changes the value of a parameter—the contact rate—within a fixed transmission mechanism, rather than moving agents to adopt a qualitatively different model of their situation. Van Basshuysen et al. (2021, 117–18) trace this mechanism: How individuals implement a given policy shapes the contact rate that feeds back into the model's own predictions. The one-step restriction, designed to regulate replacement among admissible models, therefore has no application here.

4.4. The legitimacy of the model user and the avoidance of manipulation

Within this scope, the kind of model choice considered implies a value-laden choice. Winsberg and Harvard (2024) identify three relevant positions in the debate concerning values in model choice: (i) being transparent about the values that informed a choice (Douglas 2009; Elliott 2017), (ii) using values that are ethically correct, or (iii) using values that are held by the public the model is meant to serve (Schroeder 2017; for a broader review of proposals for managing values in science, see Harvard and Winsberg 2024). According to Winsberg and Harvard (2024), (i) alone is insufficient: It lets the public detect a mismatch with its own values only if modelers can fully articulate the value commitments behind each choice—a demanding requirement for complex models. (ii) alone is also insufficient: Reasonable disagreement can persist among value commitments that are all, *ex hypothesi*, not ethically wrong, leaving a modeler free to impose whichever such values they happen to hold. (iii) is the most demanding but, on their view, the only fully satisfactory standard. However, as Winsberg and Harvard (2024) argue, this ideal is hard to reach in practice: Coproducing a model requires members of the public to grasp technical considerations that determine a model's behavior, and participatory processes risk becoming a formality that satisfies the letter of participation without its substance.

In virtue of this difficulty, a political warrant to allow the use of the principle is required. A model user's own judgment that M^+ is preferable is not by itself sufficient to license replacement. A modeler, for instance, is conscripted in virtue of expertise in constructing and evaluating representations, not in virtue of any standing to decide

what other people's ends ought to be. Were such judgment sufficient, the principle would do no more than filter which private judgments get acted upon, not whether the judgment is legitimate.

In this regard, Schroeder (2022) distinguishes ethics-grounded from politics-grounded standards for evaluating value judgments in science: Political standards give "a central place to concepts like representativeness, legitimacy, accountability, and authority" and "a greater role to procedure" (248) whereas ethical standards ask what is true or reasonable. This procedural standard is the necessary proxy: It does not require verifying that a direction matches the public's actual values, only that it derives from a process that serves as a workable substitute for that verification because direct coproduction is rarely available in practice. This blocks the reading on which the principle simply licenses whatever direction a model user happens to consider ethical: What does the normative work is the independent process, not the model user's private assessment of what is ethical.

Concretely, this means a shift toward M^+ is warranted only when the preferred direction is independently and publicly established—for instance, through legislated targets or institutional mandates. This article does not attempt to adjudicate when a given legislative or institutional process qualifies as adequately public-regarding and instead takes the existence of such a process as what distinguishes a warranted from an unwarranted direction. Public processes of this kind remain open to contestation independently of whoever invokes them.[9]

Consider the PEB case. Suppose a modeler communicates SH rather than PD in a public report on short-haul flying, framing abstinence as a matter of coordinating with others already willing to cooperate rather than as a lone sacrifice with no prospect of uptake. If the modeler does so purely on their own conviction that reducing short-haul flights is desirable, this is not licensed by the One-Step Principle, even though SH clears the same admissibility threshold as PD. If instead the modeler is operating within an established climate policy framework—one that has already set emissions targets through a legislative or otherwise representative process—the same choice is licensed because the direction it pursues is warranted independently of the modeler's own assessment.

Even when this procedural standard is met, a further difficulty remains. Thoma (2024, 23–26) argues that even a democratically warranted single privileged representation generates an epistemic inequality between citizens whose priorities it reflects and those whose it does not and recommends value-pluralistic "dashboard" representations instead (30–34). Because the One-Step Principle licenses communicating one replacement model rather than a plurality, it does not resolve this difficulty: Political warrant for the direction of replacement is not, by itself, a warrant against this inequality.

A separate concern regards the legitimacy of the model user's choices, which is closely tied to concerns about autonomy and the risk of manipulation. In this regard, the literature on nudges offers a useful framework to identify whether the use of this principle can be justified as nonmanipulative and preserving people's autonomy. In fact, like in the nudge case, this principle, by replacing a certain performative model

[9] See Lusk (2020), who argues that political legitimacy need not require full democratic ratification of every value-laden choice, provided the choice remains open to further public contestation.

with another one, alters the choice architecture without depriving people of the possibility of choosing a certain option or course of action. However, by suggesting that the representation given by a certain performative model is more likely, it aims to direct people's choices toward achieving a goal already fixed, which raises ethical challenges.

According to Schmidt and Engelen (2020), nudges can be problematic for four main reasons. First, they may undermine people's autonomy and rational agency. Second, they may lead to dominating power relationships. Third, nudging could open the door to paternalism and to policy makers imposing their own values on citizens who may reasonably disagree about what is good for them. Finally, policy makers using nudges risk manipulating people.

In response to the first objection, it could be argued that the counterfactual scenario—removing the nudge altogether—does not typically leave agents more autonomous or rational either; some nudges even improve decision making by aligning it with agents' own ends. The same holds for the principle: Even without its application, some representation of the target system will still be communicated, and this baseline is not obviously more autonomy-preserving. The second and third objections—domination and paternalism—can be met by the requirement already established in the preceding text: A shift toward M^+ is licensed only through an independently and publicly warranted process (section 4.4), so neither unchecked power nor a model user's private convictions can by themselves warrant it. The fourth objection, concerning manipulation, is addressed separately in the following text.

According to Noggle (2026), manipulation operates through one or more of three broad routes: (a) by bypassing an agent's capacity for rational deliberation, (b) by inducing a faulty belief, desire, or emotion, or (c) by exerting pressure that falls short of coercion. As for (a), see the answer to the first objection against nudging. In addition, agents are free to dissent from the view communicated and to deliberate in accordance with their rational construction of reality. As for (b), it is true that a more optimistic framing may motivate compliance partly through its affective pull, independently of any further reason to accept it as accurate. However, this is what the salience condition (see section 4.5) is meant to police: Salience requires that agents can be given reasons to recognize M^+ as credible, not merely that M^+ feels more encouraging. A replacement satisfying this condition moves agents by giving them grounds to revise their view of their situation, not by exploiting an affective response detached from those grounds. The third route is harder to rule out in general: Favorable conditions for self-fulfilling performativity (the third condition in section 4.5) may increase social pressure to conform. The principle does not eliminate this risk, but it constrains who may license it: Because the direction of replacement must derive from a public, contestable process rather than a private judgment, any pressure the process generates is at least open to challenge on the same public grounds—a safeguard the principle provides, even though it falls short of guaranteeing the absence of pressure altogether.

As Schmidt and Engelen (2020) note, nudge legitimacy is best assessed case by case. The same holds here: A complete assessment of the One-Step Principle requires analyzing specific models. Nevertheless, there are no clear knockdown arguments against its use.

4.5. The principle formalized

The One-Step Principle specifies when performative considerations may permissibly guide model selection by imposing three substantive conditions and one procedural restriction. The principle is stated here for the general class of models identified in section 4.3; PD, SH, and HL models of PEBs illustrate how each condition applies.

The first condition is perceived admissibility parity: An optimistic model may replace a more pessimistic one only if the two are in this relation, as defined in section 4.2.

The second condition is salience. I use "salience" in a broadly Schellingian sense.[10] A model is salient when it can function as a focal representation for the relevant agents: that is, when it makes a particular interpretation of the target system recognizable and action-guiding. Salience is therefore a relation between the model, the agents, and the context in which the model is communicated. Salience should not be understood as mere conformity to agents' current beliefs or to an already entrenched model. A model may be salient, or become salient, when agents can be given reasons to recognize it as a credible representation of their target system. The condition therefore allows model users to challenge existing framings, but only through representations that can plausibly become intelligible and action-guiding for the relevant agents. If a model cannot be made salient in this sense, its communication risks becoming either ineffective or manipulative: ineffective because it fails to coordinate expectations and manipulative because it asks agents to act under a representation they do not have reason to accept.

The third condition concerns favorable conditions for self-fulfilling performativity. The context must contain credible causal mechanisms through which uptake of the model can reshape expectations in ways that support the behavior the model represents. Such mechanisms may include visibility of the represented behavior, trusted communicators, institutional support, or background norms that make the relevant expectations credible. Without such mechanisms, choosing the more optimistic model would be normatively motivated but performatively inert.[11]

The procedural restriction is the one-step requirement. Even when perceived admissibility parity, salience, and favorable conditions for self-fulfilling performativity obtain, a model may be replaced only by the nearest admissible model in the normatively preferred direction. The purpose of this restriction is to reduce the risk of performative backfiring when agents are asked to revise their representation of a target system too radically. This is not to say that large representational discontinuities are epistemically objectionable as such. The point is rather pragmatic: When a model is expected to guide uptake and reshape agents' expectations, a radical shift may fail to be recognized as a credible representation of their situation. The one-step requirement should therefore be understood as a constraint on controlled representational change: It permits a shift in how the relevant strategic obstacle is represented, but only when the shift preserves enough continuity with the currently salient representation to remain credible and action-guiding.

[10] On salience, focal points, and the role of shared expectations in coordination, see Schelling (1960), Lewis (1969), and Guala (2016).

[11] It is worth noting that favorable performative conditions do not make the optimistic model epistemically superior. They become relevant once perceived admissibility parity obtains.

By a "step," I mean a minimal shift in the perceived representation of the target system. A step is not a numerical distance between models, but a relation of proximity within an ordered set of admissible representations. Such proximity depends on which features of the representation are altered and which are held fixed. In the present case, steps are defined by changes in the kind of obstacle that agents take to stand in the way of cooperation. Moving from PD to SH changes the representation from one in which defection is individually dominant to one in which cooperation is possible but dependent on mutual assurance. Moving from SH to HL changes the representation from one in which cooperation is fragile and risky to one in which agents' interests are largely aligned and the main problem is coordination on the better equilibrium. By contrast, moving directly from PD to HL would skip an intermediate change in perceived incentives and may therefore undermine the credibility or uptake on which the desired performative effect depends.

A one-step replacement is therefore the nearest admissible model in the normatively preferred direction. It modifies one feature of the target representation while preserving the other features that make the model descriptively adequate and recognizable to the relevant agents. This is why the principle is both domain-relative and generalizable. It is domain-relative because the relevant ordering depends on the specific normative elements in a given field. It is generalizable because, once such an ordering is established, the same procedural constraint applies: Do not replace a salient model with a radically more desirable one if an intermediate admissible model is available. This also means that applying the One-Step Principle requires substantive domain-specific work. In each field, one must identify the relevant family of models, the normative dimension along which their performative effects differ, and the relations among admissible alternatives.

The principle can be formulated as follows:

Let M^- and M^+ be two admissible models for a given target system, where M^+ represents the target system in a manner that is, on some normatively relevant dimension, more favorable than M^-. It is permissible to adopt or communicate M^+ in place of M^- only if:

i. M^- and M^+ are in perceived admissibility parity;
ii. M^+ is salient, or can plausibly become salient, for the relevant agents;
iii. The context contains favorable conditions for self-fulfilling performativity, such that the public uptake of M^+ could credibly reshape agents' expectations in the represented direction;
iv. M^+ differs from M^- by only one step in the ordered space of admissible models.

Applied to PD, SH, and HL models of PEBs, the decision procedure is straightforward. A shift from PD to SH is permissible only if PD and SH are perceived as being in admissibility parity, SH can plausibly become salient, and favorable conditions for self-fulfilling performativity obtain. A shift from SH to HL is permissible under the same conditions. A direct shift from PD to HL is not permissible when SH is an admissible intermediate alternative because it would violate the one-step restriction.

Where these conditions are not met, the more pessimistic model should be retained. This makes explicit the conservative character of the principle. It allows

performative considerations to guide model selection under perceived admissibility parity, when the optimistic alternative can credibly shape expectations, and only through the nearest admissible shift in the normatively preferred direction. It therefore blocks both epistemically unwarranted optimism and radical reframing that agents are unlikely to recognize as credible.

5. A conservative principle for performative model selection

The previous section made explicit the conservative character of the One-Step Principle. This section defends the principle against two risks it is designed to avoid: paternalistic steering and wishful modeling. Both risks stem from letting the desirability of an outcome override an independent constraint on model choice—legitimacy in the first case, evidence in the second. The section closes by addressing a further, conceptual challenge: whether the notion of a "step" is precise enough to generalize beyond the PD–SH–HL sequence.

The first risk is paternalistic steering: model users choosing a model because of its expected behavioral effects, thereby appearing to steer agents rather than inform them (Ortmann 2025). The One-Step Principle restricts permissible replacement to block this risk through the three conditions and the procedural restriction of its formal statement (section 4.5), and only when its use is legitimate to begin with (section 4.4). Because it applies only under perceived admissibility parity (section 4.2), it does not license communicating a model known to be inferior. Because its use presupposes legitimacy, the direction of replacement must be warranted independently of the model user's own judgment, rather than fixed by whoever happens to be selecting the model. Because it requires salience, favorable conditions for self-fulfilling performativity, and a one-step replacement, the alternative model must be one agents can recognize as credible, one whose uptake can plausibly become self-fulfilling, and the nearest such alternative rather than a radical departure from what is currently salient. These constraints narrow performative model choice to cases in which expectation-guidance can be distinguished from opportunistic steering—without pretending that such choice is ever normatively innocent.

Radical reframing remains a risk even where the one-step restriction is formally satisfied: Agents may still fail to recognize the nearest admissible model as a credible representation of their situation, undermining salience or generating resistance. In that case, the appropriate response is to alter the background conditions that would make a more optimistic representation credible, rather than to force the replacement immediately.

The second risk in performative model selection is wishful modeling: replacing a model purely because its uptake would be desirable, without checking whether the principle's conditions are met. The principle blocks this by making desirability alone insufficient: Replacement remains forbidden when the more desirable model is descriptively inadequate, cannot plausibly become salient, or lacks credible mechanisms through which its uptake could become self-fulfilling. Desirable performative effects are therefore not sufficient for model choice, addressing Winsberg and Harvard's (2022) concern that expected social benefits might override these independent, nonconsequentialist constraints.

A further objection concerns the notion of a "step": Can it generalize beyond the PD–SH–HL sequence, or is it merely a stylized feature of that example? This objection should be taken seriously. As section 4.5 establishes, a step is a relation of proximity, not a fixed metric—what the objection presses is whether that relation is precise enough to do work beyond the PEB case. The notion is therefore neither purely psychological nor merely evaluative: It is not a matter of how close two models feel to whoever is choosing between them, nor of how their expected outcomes compare, but of how the models represent the interaction. This is a criterion the bank-run case (section 4.3) already applies outside the PEB sequence, ordering its three accounts by the mechanism of withdrawal rather than by any numerical distance. The generalizability is therefore not ad hoc: It is procedural, not domain-specific—only the ordering varies across fields and identifying it in each new domain is genuine analytical work, not something for which this procedural guarantee can substitute.

6. Conclusion

This article has argued that performative model selection raises a distinctive normative problem when alternative models are perceived to be in admissibility parity but differ in their foreseeable performative effects. I developed this argument through game-theoretic models of PEBs, focusing on PD, SH, and HL as idealized representations of different strategic structures.

The central claim was that the harmful performativity of pessimistic framings does not by itself justify their replacement. Even when a PD-like model risks stabilizing pessimistic expectations, a more optimistic alternative may be selected only under specific constraints. The One-Step Principle captures these constraints: Replacement is permissible only when the alternative model is in perceived admissibility parity with the model it replaces, can plausibly become salient, is supported by favorable conditions for self-fulfilling performativity, and differs from the currently salient model by only one step. This principle applies more generally to admissible models forming a discrete, ordered family along a normatively relevant dimension. The direction of this replacement, however, must be independently and publicly warranted.

Acknowledgments. I am deeply grateful to Francesco Guala and Luca Zanetti for their guidance and feedback at every stage of this project. I thank Philippe van Basshuysen and the members of the MAPS project (Managing Performative Science) for hosting me at Leibniz University Hannover and for their helpful comments on an earlier version of this paper, in particular George Kwasi Barimah, Lukas Beck, Jakob Ortmann, Maxim Pfaffenrot, and Nadia Ruiz. My thanks also go to Edoardo Peruzzi for his generous feedback. This work was presented at the 4th Lake Como INEM Summer School, and I am grateful to the participants for their valuable suggestions, especially N. Emrah Aydinonat and Caterina Marchionni. Finally, I thank two anonymous reviewers for their insightful comments.

Funding Statement. None to declare.

Declarations. None to declare.

References

Asker, Andrea S., and H. Orri Stefánsson. 2023. "Climate Change and Decision Theory." In *Handbook of the Philosophy of Climate Change*, edited by Gianfranco Pellegrino and Marcello Di Paola, 267–86. Cham: Springer. https://doi.org/10.1007/978-3-031-07002-0_29.

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

Carrozzo Magli, Alessio, Pompeo Della Posta, and Piero Manfredi. 2021. "The Tragedy of the Commons as a Prisoner's Dilemma: Its Relevance for Sustainability Games." *Sustainability* 13 (15):8125. https://doi.org/10.3390/su13158125.

DeCanio, Stephen J., and Anders Fremstad. 2013. "Game Theory and Climate Diplomacy." *Ecological Economics* 85:177–87. https://doi.org/10.1016/j.ecolecon.2011.04.016.

Douglas, Heather E. 2009. *Science, Policy, and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press. https://doi.org/10.2307/j.ctt6wrc78.

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

Fudenberg, Drew, and Eric Maskin. 1986. "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information." *Econometrica* 54 (3):533–54. https://doi.org/10.2307/1911307.

Godman, Marion, and Caterina Marchionni. 2022. "What Should Scientists Do about (Harmful) Interactive Effects?" *European Journal for Philosophy of Science* 12 (4):63. https://doi.org/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: Princeton University Press. https://doi.org/10.1515/9780691214665-007.

Guala, Francesco. 2016. "Performativity Rationalized." In *Enacting Dismal Science: New Perspectives on the Performativity of Economics*, edited by Ivan Boldyrev and Ekaterina Svetlova, 29–52. New York: Palgrave Macmillan. https://doi.org/10.1057/978-1-137-48876-3_2.

Hardin, Garrett. 1968. "The Tragedy of the Commons." *Science* 162 (3859):1243–48. https://doi.org/10.1126/science.162.3859.1243.

Harvard, Stephanie, and Eric B. Winsberg. 2024. "'Managing Values' in Health Economics Modelling: Philosophical and Practical Considerations." *Social Science & Medicine* 358:117256. https://doi.org/10.1016/j.socscimed.2024.117256.

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

Khosrowi, Donal. 2023. "Managing Performative Models." *Philosophy of the Social Sciences* 53 (5):371–95. https://doi.org/10.1177/00483931231172455.

Kopec, Matthew. 2017. "Game Theory and the Self-Fulfilling Climate Tragedy." *Environmental Values* 26 (2):203–21. https://doi.org/10.3197/096327117X14847335385553.

Lewis, David. 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Lloyd, William Forster. 1833. *Two Lectures on the Checks to Population*. Oxford: Oxford University Press.

Lusk, Greg. 2020. "Political Legitimacy in the Democratic View: The Case of Climate Services." *Philosophy of Science* 87 (5):991–1002. https://doi.org/10.1086/710803.

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

Mailath, George J., and Larry Samuelson. 2006. *Repeated Games and Reputations: Long-Run Relationships*. New York: Oxford University Press.

Marchionni, Caterina, Julie Zahle, and Marion Godman. 2024. "Reactivity in the Human Sciences." *European Journal for Philosophy of Science* 14 (1):8. https://doi.org/10.1007/s13194-024-00571-y.

Merton, Robert K. 1948. "The Self-Fulfilling Prophecy." *Antioch Review* 8 (2):193–210. https://doi.org/10.2307/4609267.

Noggle, Robert. 2026. "The Ethics of Manipulation." In *The Stanford Encyclopedia of Philosophy*, Summer 2026 ed., edited by Edward N. Zalta and Uri Nodelman. https://plato.stanford.edu/archives/sum2026/entries/ethics-manipulation/.

Ortmann, Jakob. 2025. "Performative Paternalism." *European Journal for Philosophy of Science* 15 (2):25. https://doi.org/10.1007/s13194-025-00651-7.

Ortmann, Jakob, and Walter Veit. 2023. "Theory Roulette: Choosing That Climate Change Is Not a Tragedy of the Commons." *Environmental Values* 32 (1):65–89. https://doi.org/10.3197/096327122X16452897197784.

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

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Schmidt, Andreas T., and Bart Engelen. 2020. "The Ethics of Nudging: An Overview." *Philosophy Compass* 15 (4):e12658. https://doi.org/10.1111/phc3.12658.

Schroeder, S. Andrew. 2017. "Using Democratic Values in Science: An Objection and (Partial) Response." *Philosophy of Science* 84 (5):1044–54. https://doi.org/10.1086/694003.

Schroeder, S. Andrew. 2022. "Thinking about Values in Science: Ethical versus Political Approaches." *Canadian Journal of Philosophy* 52 (3):246–55. https://doi.org/10.1017/can.2020.41.

Thoma, Johanna. 2024. "Social Science, Policy and Democracy." *Philosophy & Public Affairs* 52 (1):5–39. https://doi.org/10.1111/papa.12250.

van Basshuysen, Philippe. 2025. "Performativity in Science: Past and Future." *Philosophy Compass* 20 (10): e70062. https://doi.org/10.1111/phc3.70062.

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):110–27. https://doi.org/10.23941/ejpe.v14i1.582.

Winsberg, Eric, and Stephanie Harvard. 2022. "Purposes and Duties in Scientific Modelling." *Journal of Epidemiology and Community Health* 76 (5):512–17. https://doi.org/10.1136/jech-2021-217666.

Winsberg, Eric, and Stephanie Harvard. 2024. *Scientific Models and Decision-Making*. Elements in the Philosophy of Science. Cambridge: Cambridge University *Press*. https://doi.org/10.1017/9781009029346.

Cite this article: Guerra, Alessandro. 2026. "The One-Step Principle: Performative Model Selection under Perceived Admissibility Parity." *Philosophy of Science*. https://doi.org/10.1017/psa.2026.10306