We Need to Talk About Self-Fulfilling Predictions[*]

Donal Khosrowi, Leibniz University Hannover, Germany
Markus Ahlers, University of Tübingen, Germany
Philippe van Basshuysen, Leibniz University Hannover, Germany

Some predictive systems do not merely predict, but their predictions shape and steer the world towards certain outcomes rather than others; they are performative. When predictive systems are performative, their development and deployment raises morally urgent challenges and places novel responsibilities on developers, deployers, regulators and policy-makers. While EWAF and other related communities have focused considerable attention on ethically significant problems regarding bias, fairness, and discrimination, little attention has been paid so far to the ethical challenges raised by performative prediction. This paper details this gap, provides a snapshot of ongoing work across computer science and philosophy to point out fruitful connections, and issues a community-wide call for action to investigate and manage performative prediction and the new challenges it raises.

Keywords: Performativity, Self-Fulfilling Prediction, Feedback Loops, Popularity Bias, Risk Assessment, Recommender Systems, Ethics, Responsibility, AI Policy.

Reference Format:
Khosrowi, Donal, Markus Ahlers, and Philippe van Basshuysen. 2025. We Need to Talk About Self-Fulfilling Predictions. In *Proceedings of the Fourth European Workshop on Algorithmic Fairness (EWAF'25)*. Proceedings of Machine Learning Research, 5 pages.

1 Introduction

The last decade has seen an explosion of work on issues of bias, discrimination, fairness, and accountability in regard to machine learning (ML) systems. This is a notable achievement of the computer science, social science, legal studies, STS and philosophy communities, attending to the moral and legal disruptions that novel technologies raise.

In recent years, a new, ethically significant challenge has drawn attention from philosophers and computer scientists, but despite its urgency, has not yet received comparable collective attention: *performative prediction*. Speaking generally, a prediction is *performative* when it causally affects the outcome that is predicted. Two types of performative prediction are often distinguished: *self-defeating* (SD) predictions, such as when an algorithmic system predicts traffic jams at junction J13, and, in response, drivers avoid routes through J13 and no traffic jam emerges; and *self-fulfilling* (SF) predictions, such as when an economic model predicts the failure of a bank and, in response to the prediction, customers move their assets from the bank, causing liquidity shortages and a subsequent bank failure.

[*] This is an extended abstract of [14], published at FAccT 2025.

Authors' Contact Information: Donal Khosrowi, Leibniz University Hannover, Hannover, Germany, donal.khosrowi@philos.uni-hannover.de; Markus Ahlers, University of Tübingen, Tübingen, Germany, markus.ahlers@uni-tuebingen.de; Philippe van Basshuysen, Leibniz University Hannover, Hannover, Germany philippe.van.basshuysen@cells.uni-hannover.de.

EWAF'25, June 30–July 02, 2025, Eindhoven, NL

SD performativity has received increasing attention by computer scientists because it poses a threat to predictive accuracy, thus implying worse outcomes for those affected by predictions and raising technical challenges. The key approach in the literature is to anticipate the second-order behavioral response of a system to a first-order prediction, and to modify an algorithmic system to account for this response in its predictions.

By contrast, SF performativity has not received much attention to date from computer scientists. It has, however, recently been increasingly explored by philosophers. SF performativity is more pernicious: while there is no threat to predictive accuracy, it can be a significant *moral* threat. For instance, risk assessment tools (RATs) such as COMPAS may issue predictions of offender recidivism risk that are viciously self-fulfilling [11]: a defendant is predicted to be at high risk to reoffend; they are subsequently incarcerated rather than released on parole; incarceration causally increases their recidivism risk; and, upon release, this elevated risk is borne out by behaviors that match the prediction. In such cases, RATs predict accurately, but for the wrong reasons: while there are no acute concerns about predictive accuracy, there are reasons to think that something is ethically problematic about algorithmic systems that make the world like they predict it to be.

The central aim of this paper is to issue a call for action to the larger ML and AI ethics communities to jointly engage SF performativity. We believe that existing computer science work on SD performativity has made important progress, but should be extended towards SF performativity, too. Unlike SD performativity, where technical aspects dominate, SF performativity is a thoroughly *epistemic-ethical problem*: it brings together technical questions about how to design algorithmic systems with performativity in mind, empirical questions about how to measure and anticipate performative effects, as well as moral, legal and regulatory questions about how to build predictive technologies responsibly that do not merely predict but have the power to significantly steer stakeholder outcomes.

2 The Landscape of Performativity

Performativity is a well-studied phenomenon. Social scientists have long investigated the epistemological problems that arise when predictions influence the expectations of the people whose behavior is being predicted, e.g. in election polling or economic forecasting [20][6][25][12]. Recent work in the philosophy of science has investigated performative scientific models [11][12][21][27][28], and highlighted that difficult normative questions arise when models steer outcomes in particular ways. In machine learning, the notion of performativity was long neglected. Perdomo et al. [23] address this issue, providing a general framework for dealing with performativity in machine learning, including solution concepts which try to account for performative effects in ML predictions. One of their key theoretical contributions is to conceptualize the performativity of an algorithmic prediction as a distribution shift caused by a predictor. Against this background, they offer two solution approaches: *performative stability* and *performative optimality*, which are constructed as optimization problems that seek to issue predictions which take into account their own performative effects in a way that minimizes the deviation between predictions and outcomes, also called *performative risk*. Perdomo et al.'s framework acknowledges that performativity implies that any prediction will steer outcomes in certain directions, so the aim is to determine which predictions steer outcomes in a way that minimizes performative risk. While their framework is the first significant attempt to deal with performativity in ML, a limitation is that it does not sufficiently acknowledge and engage the *ethical* issues that the shift from *predicting* to *steering* outcomes presents.

However, the distinction between predicting and steering is crucial: predicting is an epistemic endeavor; the world is to be described as it is or will be. But steering is, usually, a *normative* endeavor; the world is to be shaped as it should be. Beyond neglecting the normative dimensions of performativity, the existing ML literature has also only focused on SD performativity, but has yet to engage SF performativity, which is normatively more challenging. Let us outline three cases that illustrate the normative challenges arising in SF domains.

2.1 Performative Recommender Systems

Recommender systems (RS) are used to predict which things, e.g. products, news articles, social media posts, music, movies and shows, are relevant to users [17]. There is increasing awareness that RS may be performative, e.g. by predict an item to be relevant and thereby causally contributing to the item's relevance, making the prediction come true. This is often discussed under the label of 'feedback loops' [18]. RS methodology often rests on strong assumptions about user preferences: it is usually assumed that preferences are well-defined and stable over time. However, decades of research in economics and psychology suggest preferences might not be stable [3]. In a performative world, this creates additional, ethically salient worries: RS may not only fail to cater to individuals' stable preferences, but they may intervene on and change user preferences over time, including in problematic ways [4][18][21]. For instance, streaming platform RS are suspected to fall prey to gender-correlated bias in item ratings because males tend to rate and review items more frequently than females. This could lead female users to engage more frequently with male-preferred shows over time and change their substantive preferences. On social platforms such as YouTube, where users also inform themselves politically, RS can contribute to political radicalization of users [9][24][30]. The performativity of RS in the form of feedback loops and popularity bias is one of the explanations for this phenomenon: RS do not only predict item relevance but can plausibly *make* contents relevant, even though the public may not consider them relevant otherwise.

2.2 Performative Risk Assessment Tools

ML-based risk assessment tools (RATs) are used for recidivism risk prediction in criminal justice [10]. As suggested earlier, an important concern is that RATs like COMPAS may not only predict recidivism but causally affect subjects' propensity to reoffend [11]. A defendant classified as high risk may be incarcerated rather than released on parole, and incarceration itself may causally contribute to elevated reoffense probabilities upon release, making the prediction come true. An increasing amount of empirical evidence suggests that incarceration increases recidivism by disrupting offenders' family ties and friendships, leading to loss of employment and difficulties finding employment and housing upon reentry into society [1][2][26][28]. Individuals' propensity to reoffend is not a static property but is an endogenous quantity that depends on judicial decisions and the predictions that inform them. Current methods to develop RATs, however, do not account for performativity [11]. Their classification accuracy is evaluated using observational data, which cannot tell us what a particular defendants' probability to reoffend would be under different treatments, such as incarceration vs. release on parole. RATs hence carry the risk of incarcerating putatively high-risk offenders whose recidivism probability would have been lower under an alternative sentence. Once again, SF performativity invites significant ethical challenges and may require rethinking of how predictive tools function.

2.3 ML in Scientific Discovery

There has recently been increasing interest in building ML systems to predict upcoming scientific discoveries and suggest fruitful ideas, hypotheses or experiments to researchers [16][7][8]. As argued in [13] these *predictive discovery and recommender systems* (PDRS) raise concerns about performativity. Scientists may respond to predictions and recommendations made by PDRS, either by pursuing them (SF) based on compelling predictive track records or by strategically avoiding them (SD) to secure priority on discoveries not anticipated by PDRS. What is more, PDRS trained on existing scientific literature may also perpetuate undesirable features of the scientific enterprise [19], such as incentives for researchers to produce marginal discoveries for novelty's sake, or by crowding out diversity in ideas and perspectives through homogenous recommendations. As [13] notes, current PDRS do not sufficiently distinguish between *predicting* discoveries that are likely and *recommending* discoveries that are valuable or significant. However, recommendation and steering, unlike mere prediction, are a normative endeavor: where PDRS should steer scientific discovery trajectories is a value-laden question that technical work alone cannot address.

2.4 Ethical Topology

Across the cases we present here, we can make some general observations to contour the landscape of SF prediction more sharply. First, not all SF cases are the same: some predictions affect decision-subjects without an active behavioral response on their part, such as in recidivism risk prediction, where defendants do not *choose* their incarceration outcomes or how imprisonment affects their propensity for criminal behaviors downstream. Other predictions act on outcomes while decision-makers have more agency, e.g. to engage recommended social media contents, or follow a recommendation for a discovery project, and sometimes their agency might be promoted (e.g. alerting researchers to discovery trajectories that they would have not considered otherwise), whereas other times it might be diminished (trapping social platform users in distributions of contents that blind them to other relevant information). Such differences are ethically significant and require particular attention.

Second, saying that SF performativity is an irreducibly ethical affair doesn’t imply that there are always *bad* things happening when it arises. For instance, in medical contexts, an SF prediction might inform a patient about what likely health outcomes they may expect if they exercised more, drank less and gave up smoking. In such cases, predictions can be self-fulfilling, yet empowering and beneficial to subjects. However, just like thinking about costs and harms to subjects must involve ethical reflection, considering and weighting the potential benefits of SF predictions, too, must involve value-judgments.

Third, while we insist that SF performativity is necessarily ethically laden, different cases exhibit varying ethical urgency. While underdetermined choices about where to steer outcomes through predictions must necessarily involve some value-judgment of what is good and bad, this does not always raise elevated ethical *concerns*: when the stakes for prediction-subjects are relatively low, e.g. predicting how busy pubs are likely to be at certain times, there are still value-judgments involved, but perhaps this doesn’t require the same attention as other, more urgent cases. Nevertheless, we should able to articulate when and why SF prediction does not require attention.

3 Managing Performativity is a Team Effort

SF performativity is an epistemically and ethically significant phenomenon and the heavily predictive character of machine learning approaches invites this phenomenon to occur across a broad spectrum of domains. However, as our cases demonstrate, the existing literature falls short in two ways: it hasn't paid sufficient attention to SF performativity, failing to recognize that ostensibly predictive ML systems do not always *merely* predict but may also causally shape the world through their predictions. Second, the literature to date hasn't sufficiently recognized that SF performativity is an irreducibly *epistemic-ethical* problem. Future work must hence involve more thorough reflection on the values necessarily involved in performative prediction. Here, we outline a three-pillar research program that engages the technical *and* ethical challenges performativity poses:

Pillar 1 aims at developing methods and tools to understand, diagnose and manage performativity. This involves four sub-aims: 1) developing general theoretical frameworks that provide a comprehensive characterization of performativity, unifying SD and SF variants as well as drawing out important differences, 2) developing empirical methods to estimate performative effects using causal inference frameworks, 3) creating simulation-based frameworks to systematically study how performative effects may manifest and propagate, and 4) developing auditing protocols to examine predictive systems for performativity.

Pillar 2 involves foundational ethics research that contours the ethical issues that performativity raises. The cases discussed here suggest that tradeoffs exist between different values. For instance, predictive accuracy may stand in tension with authentic preference formation or user agency in the case of recommender systems; avoidable harms to defendants in the case of RATs, or diversity and creativity of ideas and approaches in scientific discovery settings. Foundational ethics work should facilitate broader societal discourse around how we may resolve such tradeoffs.

Pillar 3 aims at building legal and regulatory frameworks to allocate responsibility for performative effects. This involves bringing together philosophers, legal scholars, regulators, policy-makers and stakeholders alongside developers to conceive of guidelines for the responsible development of performative predictive systems. Key questions include: Who is responsible for managing performative effects? And what general principles should guide their efforts?

For instance, large social platforms employing RS hold significant powers in shaping real-world outcomes, affecting not only individuals but society at large. Yet, while potentially bearing some as of yet underappreciated moral responsibility for harmful performative effects, developers and deployers are also ill-positioned to legitimately make value-laden choices on behalf of society, determining *where* to steer outcomes, if performativity is indeed inevitable. Policy-makers and regulators, who *are* suitably legitimized, must hence be sensitized to the urgency of the problems performativity raises and deploy regulatory and policy tools to manage performativity. So far, emerging regulation like the EU AI Act is not prepared to recognize performativity as a distinct phenomenon, potentially requiring specific measures such as performativity audits.

4 Conclusions

Some predictive systems do not merely predict, but their predictions steer the world towards certain outcomes rather than others: they are *performative*. When predictive systems are performative, their development and deployment raises morally urgent challenges, and may place novel responsibilities on developers, deployers, regulators and policy-makers. This paper highlighted gaps in existing work, provided a snapshot of ongoing work across computer science and philosophy, and issued a community-wide call for action to investigate and manage performative prediction. We have sketched a rough, three-pillar plan to outline our vision of the next wave of research and policy on performativity and hope that the call, overview and arguments presented here will invite researchers from across the community to tackle performativity out in the open, and with the joint aim of making predictive systems perform well, and responsibly.

Acknowledgements

We wish to thank three anonymous reviewers and the area chair for their valuable comments and suggestions on our manuscript. Philippe van Basshuysen's contribution is funded by the European Union (ERC, MAPS, 101115973). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European

Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. Markus Ahlers is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2046/1 – Project number 390727645.

References

[1] Cid, José. 2009. Is Imprisonment Criminogenic? A Comparative Study of Recidivism Rates between Prison and Suspended Prison Sanctions. *European Journal of Criminology* 6(6):459-480. http://doi.org/10.1177/1477370809341128

[2] Clear, Todd R. 2008. The Effects of High Imprisonment Rates on Communities. *Crime And Justice*, 37(1), 97–132. https://doi.org/10.1086/522360

[3] Fehr, Ernst, and Karla Hoff. 2011. Introduction: Tastes, Castes and Culture: The Influence of Society on Preferences. *The Economic Journal*, 121(556): F396–F412. https://doi.org/10.1111/j.1468-0297.2011.02478.x

[4] Fletcher, Amelia, Peter L. Ormosi, and Rahul Savani. 2023. Recommender Systems and Supplier Competition on Platforms. *Journal of Competition Law & Economics*, 2023, 19, 397–426 https://doi.org/10.1093/joclec/nhad009

[5] Godman, Marion, and Caterina Marchionni. 2022. What should scientists do about (harmful) interactive effects? *European Journal for Philosophy Of Science*, 12(4). https://doi.org/10.1007/s13194-022-00493-7

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

[7] Gu, Xuemei, and Mario Krenn. 2024a. Interesting Scientific Idea Generation Using Knowledge Graphs and LLMs: Evaluations with 100 Research Group Leaders. arXiv.2405.17044

[8] Gu, Xuemei, and Mario Krenn. 2024b. Forecasting high-impact research topics via machine learning on evolving knowledge graphs. arXiv:2402.08640.

[9] Kaiser, Jonas, and Adrian Rauchfleisch. 2019. Integrating Concepts of Counterpublics into Generalised Public Sphere Frameworks: Contemporary Transformations in Radical Forms. *Javnost - The Public*, 26(3), 241–257. https://doi.org/10.1080/13183222.2018.1558676

[10] Kehl, Danielle, Priscilla Guo, and Samuel Kessler. 2017*. Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing. Responsive Communities Initiative*, Berkman Klein Center for Internet & Society, Harvard Law School.

[11] Khosrowi, Donal, and Philippe van Basshuysen. 2024. Making a Murderer – How risk assessment tools may produce rather than predict criminal behavior. *American Philosophical Quarterly* (2024) 61 (4): 309–325. https://doi.org/10.5406/21521123.61.4.02

[12] Khosrowi, Donal. 2023. Managing Performative Models. *Philosophy of the Social Sciences* 53(5):371-395.

[13] Khosrowi, Donal. 2025. Automating Pursuitworthiness: Four Concerns About the Proper Roles for Machine Learning Systems in Scientific Discovery. Preprint available at: https://seafile.cloud.uni-hannover.de/f/ee1e6c60264d4edf8d5f/?dl=1

[14] Khosrowi, Donal, Markus Ahlers, and Philippe van Basshuysen. 2025. When Predictions are More Than Predictions: Self-Fulfilling Performativity and the Road Towards Morally Responsible Predictive Systems. In *The 2025 ACM*

Conference on Fairness, Accountability, and Transparency (FAccT ’25), June 23–26, 2025, Athens, Greece. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3715275.3732072

[15] Koskinen, Inkeri. 2022. Reactivity as a tool in emancipatory activist research. *European Journal for Philosophy of Science* 12:65. https://doi.org/10.1007/s13194-022-00487-5

[16] Krenn, Mario, Lorenzo Buffoni, Bruno Coutinho, et al. 2023. Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network. *Nature Machine Intelligence*, 5: 1326-1335. https://doi.org/10.1038/s42256-023-00735-0

[17] Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

[18] Mansoury, Masoud, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke 2020. Feedback loop and bias amplification in recommender systems. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 2145–48.

[19] Messeri, Lisa, and Molly J. Crockett. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature* 627(8002):49-58. doi: 10.1038/s41586-024-07146-0.

[20] Morgenstern, Oskar. 1928. *Wirtschaftsprognose: Eine Untersuchung ihrer Voraussetzungen und Möglichkeiten*. Berlin & Heidelberg: Springer. https://doi.org/10.1007/978-3-7091-2157-3

[21] Ortmann, Jakob. 2025. Performative paternalism. *European Journal for Philosophy of Science* 15, 25. https://doi.org/10.1007/s13194-025-00651-7

[22] Pajkovic, Niko. 2022. Algorithms and taste-making: Exposing the Netflix Recommender System’s operational logics. *Convergence*, 28(1), 214-235. https://doi.org/10.1177/13548565211014464

[23] Perdomo, Juan C., Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. Performative Prediction. In *International Conference on Machine Learning (ICML)*, volume 119, pages 7599-7609. PMLR

[24] Ribeiro, Manoel Horta, Raphael Ottoni, Robert West, Virgilio A. F. Almeida, and Wagner Meira, Jr. 2019. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* ’20)*. ACM. https://doi.org/10.1145/3351095.3372879

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

[26] Spohn, Cassia, and David Holleran. 2002. The Effect of Imprisonment on Recidivism Rates of Felony Offenders: a Focus on Drug Offenders. *Criminology*, 40(2), 329–358. https://doi.org/10.1111/j.1745-9125.2002.tb00959.x

[27] van Basshuysen, Philippe. 2023. Austinian Model Evaluation. *Philosophy of Science* 90 (5): 1459–68. doi:10.1017/psa.2023.24.

[28] 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). doi:10.23941/ejpe.v14i1.582.

[29] Vieraitis, Lynne M., Tomislav V. Kovandzic, and Thomas B. Marvell. 2007. The criminogenic effects of imprisonment: Evidence from state panel data, 1974–2002, *Criminology and Public Policy*, vol. 6, no. 3, pp. 589-622.

[30] Whittaker, J. & Looney, S. & Reed, A. & Votta, F. 2021. Recommender systems and the amplification of extremist content. *Internet Policy Review*, 10(2). https://doi.org/10.14763/2021.2.1565