HHS Public Access

Author manuscript

Resuscitation. Author manuscript; available in PMC 2024 February 01.

Published in final edited form as:

Resuscitation. 2023 February ; 183: 109622. doi:10.1016/j.resuscitation.2022.10.014.

Self-fulfilling prophecies and machine learning in resuscitation science

Maria De-Arteaga, PhD[1], **Jonathan Elmer, MD, MS**[2]

[1]Information, Risk and Operations Management Department, McCombs School of Business, University of Texas at Austin, Austin, TX USA

[2]Departments of Emergency Medicine, Critical Care Medicine and Neurology, University of Pittsburgh School of Medicine, Pittsburgh, PA USA

Abstract

Introduction: Growth of machine learning (ML) in healthcare has increased potential for observational data to guide clinical practice systematically. This can create self-fulfilling prophecies (SFPs), which arise when prediction of an outcome increases the chance that the outcome occurs.

Methods: We performed a scoping review, searching PubMed and ArXiv using terms related to machine learning, algorithmic fairness and bias. We reviewed results and selected manuscripts for inclusion based on expert opinion of well-designed or key studies and review articles. We summarized these articles to explore how use of ML can create, perpetuate or compound SFPs, and offer recommendations to mitigate these risks.

Results: We identify four key mechanisms through which SFPs may be reproduced or compounded by ML. First, imperfect human beliefs and behavior may be encoded as SFPs when treatment decisions are not accounted for. Since patient outcomes are influenced by a myriad of clinical actions, many of which are not collected in data, this is common. Second, human-machine interaction may compound SFPs through a cycle of mutual reinforcement. Third, ML may introduce new SFPs stemming from incorrect predictions. Finally, historically correct clinical choices may become SFPs in the face of medical progress.

Conclusion: There is a need for broad recognition of SFPs as ML is increasingly applied in resuscitation science and across medicine. Acknowledging this challenge is crucial to inform research and practice that can transform ML from a tool that risks obfuscating and compounding SFPs into one that sheds light on and mitigates SFPs.

Keywords

Machine learning; prediction; outcome; bias

Corresponding author: Jonathan Elmer, MD, MS, Iroquois Building, Suite 400A, 3600 Forbes Avenue, Pittsburgh, PA 15206, elmerjp@upmc.edu.

Author's contributions: Both authors conceived of and drafted the manuscript, provided critical revisions and approved the final manuscript. MD-A prepared the figures.

Competing interest: None.

Introduction

Humanity has been preoccupied with causes and consequences of self-fulfilling prophecies (SFPs) for millennia.[1–3] In medicine, fear of SFPs shapes our approach to prognostication and resulting treatment decisions, particularly in settings where life-sustaining therapies may be withdrawn or withheld based on perceived poor prognosis.[4] SFPs affect the foundation on which biomedical knowledge is built – learning through observation of clinical outcomes – so can be perpetuated under the veil of evidence-based practice. This risk is exacerbated by growth in machine learning (ML), which has rapidly increased available tools that allow learning from observational data.

A common goal is to estimate the probability of a future event or outcome (Y) given a set of clinical information (X). In reality, the observed outcome Y does not depend on X alone. Observed outcomes are also influenced by myriad treatment decisions that in turn are affected by providers' outcome predictions, which may also affect X. In settings where shared decision-making occurs, Y may also be influenced by patient or surrogate perception of the anticipated outcome. The type of data used for training, flaws in modeling, human-machine interaction, and changes in medical knowledge may all lead to SFPs. While growth in ML has been accompanied by rigorous consideration of many sources of biases,[5–7] the problem of SFPs has received relatively little attention.

We define a SFP as a self-perpetuating or self-amplifying source of bias through which prediction of an outcome increases the chance the outcome will occur.[4] It should be noted that the perils of SFPs in medicine are not limited to quantitative analysis. Cognitive heuristics (i.e., clinical gestalt), implicit biases, and other factors that alter treatment decisions also create SFPs. Guarding against these perils requires careful identification of the putative mechanism(s) at play, making it imperative to understand how this long-standing concern in healthcare interplays with new technologies. We dissect how the deployment of ML systems may create, perpetuate or compound SFPs, and offer recommendations to mitigate this risk.

Methods

We conducted a scoping review that aimed to identify distinct themes relevant to SFPs and ML in medicine and resuscitation science. We searched PubMed and ArXiv for English language articles published through January 2021. We used the search terms: ("machine learning" OR "artificial intelligence") AND ("algorithmic fairness" OR "bias" OR "self-fulfilling prophecies"). For arXiv, we constrained to include ("healthcare" OR "medicine"). The authors screened titles and abstracts of the results then reviewed the full text of potentially relevant articles. We selected articles for inclusion based on expert opinion. We supplemented these articles with several well-known examples of SFPs from the literature. We summarized key themes in a narrative review. We performed a series of illustrative simulations, the methods for which are described below.

Search Results

Our search identified 1,327 articles on PubMed and 113 articles on arXiv. We selected 29 articles for inclusion to which we added 11 articles discussing well-known examples of SFPs. We synthesized our results to identify four key mechanisms through which SFPs may be reproduced or compounded by ML (Figure 1). First, imperfect prior clinical beliefs and behavior may be encoded as SFPs when resulting treatment decisions are not accounted for. Since outcomes are influenced by innumerable clinical actions, many of which are not measured, this is common. Second, human-machine interaction may compound SFPs through a cycle of mutual reinforcement. Third, ML may introduce new SFPs stemming from incorrect algorithmic predictions. Finally, historically correct clinical choices and resulting outcomes may become SFPs in the face of medical progress.

Human self-fulfilling prophecies encoded in algorithmic predictions

Awareness that clinicians' beliefs influence outcomes has shaped medical research in fundamental ways. A clear example is recognition of the importance of double blinding in randomized trials, such that providers' knowledge of a subject's treatment allocation cannot influence study results.[8] Well-blinded, randomized trials are a gold-standard for evidence generation but are often infeasible. By contrast, observational data are widely available in the digital age and predictive models trained on these data may be comparatively easier to develop.[9,10] Studies of, and technologies trained using, observational data lack rigorous control of contextual factors and confounders.[11,12]

SFP may result when ML is used to inform clinical decisions if clinicians' prior beliefs and behaviors are encoded into model predictions. Consider a model to predict the outcome of comatose patients early after cardiac arrest. Transferring these patients to specialty care may improve outcomes.[13,14] Providers may be less likely to transfer patients they sincerely believe have little chance of favorable recovery. Providers are not omniscient and thus have imperfect and potentially systematically biased estimates of recovery potential. Average outcomes of non-transferred patients will be worsened based on this treatment decision. Models (and providers) trained to predict outcomes based on data available prior to transfer may learn erroneous relationships between clinical patterns that predict the *decision* not to transfer and the likelihood of poor outcomes.[1] Subtle factors also influence outcomes, such as the intensity of nursing care, where optimism, clinical concern, and other intangibles may motivate more (or less) attentive care, introducing biases that are pervasive and difficult to detect. Using models trained on datasets contaminated by these factors to inform future policies can perpetuate SFP, potentially preventing patients for whom outcomes historically were predicted to be poor from receiving care that could improve chances of recovery. While in the case of transfer, it may be relatively easy to account for this treatment choice, outcomes are affected by many clinical decisions that are never recorded.

The risks of ML learning spurious correlations arising from treatment decisions have been considered, but the ways this can perpetuate SFPs is under-studied. Existing literature has

[1]Technically, the problem can be formalized as there being unobserved mediators, Z, so that the probability of Y depends on both X and Z, but the model is unable to control for the treatment effects, Z.

often focused on examples in which patient factors that drive best medical practices result in SFPs that yield better-than-expected patient outcomes. Less attention has been devoted to cases in which treatment choices negatively impact outcomes. Crucially, solutions that address the former do not necessarily address the latter. Consider a well-known example of an ML system designed to predict risk of death from pneumonia.[15] Researchers found an apparent protective effect of asthma resulting in lower predicted probability of death among asthmatics in the training data, an observation that defies clinical intuition. This is explained by particularly attentive care asthmatic patients received, which improved their outcomes. In short, asthma was a marker of better care, not a biological factor resulting in lower risk-adjusted mortality.

As experts have considered this classic example, they have proposed solutions. We highlight several commonly applied to address spurious correlations but *do not* mitigate the risk of SFPs that arise from treatment decisions. One potential solution is to use interpretable models that are refined by experts to incorporate prior knowledge, e.g. asthma being a predictor of high risk.[15] Others have argued that outcome estimates biased by treatment effects are not necessarily problematic, as long as experts use them to complement their judgment and practices.[16] These solutions fail when model predictions are mediated by poor medical choices (rather than best practice) because models learn to replicate mistakes, and experts are not necessarily well-positioned to complement or correct the algorithm. Consider the use of risk assessment systems to determine whether a patient's chance of survival is high enough to justify transfer to specialty care after severe acute brain injury, when perceived poor neurological prognosis can lead to withdrawal of life-sustaining therapies.[17,18] If suboptimal medical practices or medical nihilism result in low chances of positive outcomes for a subset of patients,[19,20] this subset could be estimated to have a chance of recovery below a threshold that warrants expensive or invasive treatments, thus perpetuating a SFP that prevents these patients from receiving care that could improve their recovery odds. In such cases, providing experts with the opportunity to complement the algorithm with their knowledge would not suffice, because errors in their own assessments would not allow them to identify these as mistaken predictions.

Human-machine interaction and compounding self-fulfilling prophecies

There are many reasons why ML algorithms might be developed and used as decision support tools rather than autonomous agents, retaining clinicians as the ultimate decision makers.[21] How clinicians integrate algorithmic predictions is complex.[22–26] Treatment recommendations interact with cognitive heuristics, biases such as optimism or nihilism, and myriad other factors.[27] The results are hard to anticipate. A human-in-the-loop offers some advantages, but human-ML collaboration may not just perpetuate but also *amplify* SFPs over time.

Consider a model predicting return of spontaneous circulation (ROSC) after cardiac arrest based on duration of cardiopulmonary resuscitation (CPR). This one-dimensional example allows us to illustrate what may happen in more complex models, while retaining the intuition afforded from a toy example. Not all patients who suffer sudden cardiac arrest achieve ROSC. Over time, the probability of ROSC falls despite ongoing CPR and

eventually precludes ROSC. Patients without ROSC do not undergo perpetual CPR–instead, providers eventually choose termination of resuscitation (TOR). It is likely providers choose TOR when they sincerely believe continuing CPR is futile (i.e., the probability of observing future ROSC if CPR were continued is believed to be extremely low). For an individual, the probability of ROSC drops to zero after TOR. In other words, TOR is deterministic of poor outcome. Past observational research suggests <5% of patients who remain pulseless after 20 minutes of CPR will achieve ROSC with continued resuscitation, few of whom enjoy favorable long-term outcomes.[28] Thus, evidence-based treatment guidelines might recommend TOR after 20 minutes.[29]

Now, assume that the authors of these guidelines have access to prospective data and annually update the recommended maximum duration of CPR based on new evidence, with the reasonable goal of preventing futile resuscitative efforts while accounting for possible improvements and changes in emergency medical service (EMS) practices. In simulation, we consider how SFPs may be perpetuated or amplified if observational data are used iteratively to estimate the CPR duration after which probability of recovery with continued CPR drops below 5%.[2] We study two scenarios (Supplemental Appendix 1). First, imagine EMS providers are perfectly adherent with guideline-recommended TOR, with only trivial random variability in actual CPR duration. Second, assume providers may approach CPR with unconscious tendencies that result in shorter or less aggressive resuscitative attempts in the two minutes preceding guideline-recommended TOR. These tendencies could stem from therapeutic nihilism, frustration for what appears to be a failing resuscitative effort, or other human factors.

Under both scenarios, thousands of excess deaths would occur annually, a result that is drastically amplified if guidelines induce or interact with human decision makers' unconscious attitudes (Figure 2). Excess deaths assuming perfect adherence to guidelines stem from the fact that those patients who would achieve ROSC after 20 minutes no longer do so. When the algorithm interacts with human biases, updated guidelines progressively shorten recommended CPR duration, as illustrated in Figure 2.

Figure 2 illustrates the risks of misestimation associated with learning from observational data in dynamic settings in which ML recommendations impact treatment choices. Broadly speaking, when models are affected by past decisions, this may not only perpetuate SFPs, but may also *compound* and *aggravate* SFPs by obscuring the link between past treatment choices and the models' predictions. Because ML may be perceived as an advanced approach, its use may provide a veneer of rigor despite persistence of the inherent limitations of observational studies.

Mistaken predictions and new self-fulfilling prophecies

Predictive algorithms may create *new* SFPs when incorrect predictions alter clinical decisions and outcomes. Algorithmic errors may stem from analytical issues such as poor choice of models, overfitting, or poor data quality, but also from flawed problem

[2]In simple notation, we consider outcome (ROSC) to be Y, the treatment decision for TOR to be D, and CPR duration to be X. Thus, the model is trained to make a treatment recommendation when $P(Y=1|X, D=1) < 0.05$.

formulation, such as choosing to optimize for a proxy that does match the outcome of interest. While clinicians are often inconsistent in their mistakes, prediction errors in ML are systematic. This consistency means that mistaken algorithmic predictions easily give rise to SFPs. Models often seek to learn continually from new information. While regular retraining is generally recommended as a best practice to account for drifts in data distribution, it allows errors both to result from and be propagated by spurious algorithmic recommendations. Consider an algorithm to estimate recovery potential in critical illness that initially underestimates the probability of recovery for patients with a particular characteristic. If these miscalibrated estimates alter treatment, such as reducing the likelihood of prescribing high-cost, invasive therapies because of the anticipation of poor prognosis, a new SFP may be introduced and perpetuated as the algorithm learns over time. The model performance estimated on observational data will not identify these as mistaken predictions and may even misleadingly indicate an improvement of predictive power over time.

The risk of introducing new SFPs through use of predictive models should be considered in conjunction with the literature on algorithmic bias,[6,7] which highlights that prediction errors often disproportionately affect historically underserved populations. Consider a study showing an algorithm used to identify patients to whom additional healthcare resources should be directed systematically underestimated the needs of Black patients.[30] The observed bias in this case resulted from use of claims-based data as a proxy for need. Historically, Black patients have had less access to healthcare than their white counterparts, and thus generate fewer insurance claims. If this algorithm were used to inform which patients require additional care, the result would be a SFP in which patients who are predicted to have fewer needs receive less care. This in turn would be reflected in low health care spending and misinterpreted as a confirmation of low health needs. As in the CPR example above, if algorithmic predictions are used to inform future practice and retrained prospectively, the magnitude of the SFP grows over time.

Stagnant predictions in the face of medical progress

Predictive algorithms exploit patterns in observational data, so assume past associations remain relevant in the future. In medicine, the stability of observed associations over time is limited by the fact that treatments and outcomes often improve. Breakthroughs in available therapeutics or treatment strategies can rapidly change the association of presenting illness severity and outcome. For example, outcomes from acute ischemic stroke were revolutionized first by widespread adoption of systemic thrombolysis and thereafter by endovascular interventions. Ongoing research has open new treatments and redefined the therapeutic window from symptom onset to intervention.[31–33] A predictive algorithm trained on historical data would systematically and falsely predict poor outcomes for patients with delayed presentation of a large vessel occlusion and might perpetuate inappropriate therapeutic nihilism.

In other cases, secular trends are more subtle. Despite the absence of efficacious new evidence-based interventions, mortality after aneurysmal subarachnoid hemorrhage, sepsis (and many other conditions) is decreasing steadily over time.[34] Unless historical trends are

projected into the future (an approach fraught with imprecision), well-performing models become inaccurate in the setting of progress. Insofar as algorithmic predictions guide care, this can introduce SFPs that may stifle medical progress.

Regularly retraining a model is often proposed as a solution to mitigate this challenge.[35] This, however, will only prevent SFPs when experts have the discretion to deviate from algorithmic recommendations, generating new data from which the algorithm can relearn associations reflected in novel treatments. At this point, tensions between mitigating this and other types of SFPs arise, since as noted above some SFPs may be amplified when retraining.

Paths forward

Providing recommendations for best practices in the face of SFPs, Wilkinson argues it is "imperative that doctors are honest with themselves and with patients and their families about uncertainty and the limits of knowledge."[4] In their current form, many ML applications in healthcare aim to leverage all available correlations with the goal of providing confident predictions, thereby obfuscating uncertainty. This is *not* an inherent property of statistical prediction, as uncertainty estimation has long been a central focus of statistics research. Appreciation of the value of uncertainty could transform ML from a tool that obscures it into one that helps clearly communicate it.[36,37] This will require research aimed at accounting for the effect predictions may have on future outcomes,[38] properly estimating prediction uncertainty associated with treatment effects,[39,40] and effective ways of communicating uncertainty to providers.[36,41]

A 2019 Scientific Statement from the American Heart Association outlines several considerations pertinent to SFPs, focusing on post-arrest neurological prognostication and subsequent withdrawal of life-sustaining therapies, which are broadly relevant beyond this specific task.[42] Training models on data accrued in settings with withdrawal of life-sustaining therapies is prohibited or strictly protocolized and potential confounders are minimized can help prevent providers' biases, mistaken beliefs and clinical choices from becoming encoded in algorithmic predictions. This approach has been used to quantify predictive value of several prognostic tests after cardiac arrest,[43] but has been less consistently used in the ML literature. An unanswered question is the extent to which other aspects of care in these cohorts affect transfer of trained models to more general settings. A second recommendation is that providers be blinded to predictions or test results during prospective evaluation of novel tools. This minimizes the potential for human-machine interactions to compound SFPs but is impractical outside of research insofar as many predictive algorithms are developed with the specific goal of informing clinical decision-making.

Human-interpretable ML also plays an important role.[44,45] Such approaches can provide information that allows providers to identify when historical treatment choices may inappropriately influence a prediction, or when novel treatments may not be accounted for. Designing tools that allow experts to actively interrogate ML recommendations is key to enabling integration of algorithmic recommendations into clinically justified decisions.[26] Here, a crucial, open question is how to design explanations and contextual settings that

mitigate rather than *exacerbate* overreliance and SFPs. While in some settings explanations have been found to increase unwarranted reliance on ML predictions,[46–48] explanations that encourage productive second guessing could reduce the risks of SFPs.

Other forms of transparency, which require less technical solutions, are also important. "Black box" properties of ML are not limited to the way in which an algorithm makes inferences, but also include users' understandings of what information is and is not available to the algorithm, what type of data were used to train the algorithm, and what target label is being predicted. In current practice, this information is often unavailable, but clarity over these questions could help providers complement algorithmic predictions with domain knowledge. An line of research on human-computer interaction and organizational sciences seeks to understand how humans integrate ML recommendations into decisions, what type of cognitive cues reduce algorithm aversion and automation bias, and how different types of training impact this.[24,25,49] Grounded on this work, there is a need for guidelines, frameworks and eventual regulations to clarify (1) what information must be clearly communicated by system designers, (2) how should this information be conveyed, and (3) what type of instruction and training I sneeded before providers start using a system.

Conclusion

There is an urgent need to recognize the risk of self-fulfilling prophecies as ML is integrated into clinical decision making. We have characterized key pathways through which SFPs may be replicated, compounded, and introduced by predictive models used for decision support.

Acknowledging the challenge posed by SFPs can open the gate to research and practice that transforms ML from a tool that risks obfuscating and compounding SFPs into one that sheds light on and mitigates SFPs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding:

Dr. Elmer's research time is supported by the NIH through grant 5K23NS097629.

Availability of data and materials:

Reproducible code for simulation analysis is included in the Supplemental Appendix.

References

1. Grene D, Aeschylus, Sophocles, Euripides. Three Greek tragedies in translation. Chicago, Ill.,: The University of Chicago press; 1942.
2. Merton RK. The Self-Fulfilling Prophecy. The Antioch Review. 1948;8(2):193–210.
3. Smith JD. The Mahābhārata. New Delhi: Penguin; 2009.
4. Wilkinson D The self-fulfilling prophecy in intensive care. Theor Med Bioeth. 2009;30(6):401–410. [PubMed: 19943193]

5. Chen IY, Pierson E, Rose S, Joshi S, Ferryman K, Ghassemi M. Ethical Machine Learning in Healthcare. Annual Review of Biomedical Data Science. 2021;4(1):null.
6. Barocas S, Selbst AD. Big Data's Disparate Impact. California Law Review. 2016;104(3):671–732.
7. Mitchell S, Potash E, Barocas S, D'Amour A, Lum K. Algorithmic Fairness: Choices, Assumptions, and Definitions. Annual Review of Statistics and Its Application. 2021;8(1):141–163.
8. Day SJ, Altman DG. Blinding in clinical trials and other studies. BMJ. 2000;321(7259):504. [PubMed: 10948038]
9. Callahan A, Shah NH. Chapter 19 - Machine Learning in Healthcare. In: Sheikh A, Cresswell KM, Wright A, Bates DW, eds. Key Advances in Clinical Informatics. Academic Press; 2017:279–291.
10. Wiens J, Shenoy ES. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. Clin Infect Dis. 2018;66(1):149–153. [PubMed: 29020316]
11. Research and Reporting Considerations for Observational Studies Using Electronic Health Record Data. Annals of Internal Medicine. 2020;172(11_Supplement):S79–S84. [PubMed: 32479175]
12. Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-World Evidence - What Is It and What Can It Tell Us? N Engl J Med. 2016;375(23):2293–2297. [PubMed: 27959688]
13. Berg KM, Cheng A, Panchal AR, et al. Part 7: Systems of Care: 2020 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. Circulation. 2020;142(16_suppl_2):S580–S604. [PubMed: 33081524]
14. Elmer J, Callaway CW, Chang CH, et al. Long-Term Outcomes of Out-of-Hospital Cardiac Arrest Care at Regionalized Centers. Ann Emerg Med. 2019;73(1):29–39. [PubMed: 30060961]
15. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery; 2015:1721–1730.
16. London AJ. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. Hastings Cent Rep. 2019;49(1):15–21.
17. Mertens M, King OC, van Putten M, Boenink M. Can we learn from hidden mistakes? Self-fulfilling prophecy and responsible neuroprognostic innovation. J Med Ethics. 2021.
18. McCracken DJ, Lovasik BP, McCracken CE, et al. The Intracerebral Hemorrhage Score: A Self-Fulfilling Prophecy? Neurosurgery. 2019;84(3):741–748. [PubMed: 29762777]
19. Hemphill JC 3rd, Newman J, Zhao S, Johnston SC. Hospital usage of early do-not-resuscitate orders and outcome after intracerebral hemorrhage. Stroke. 2004;35(5):1130–1134. [PubMed: 15044768]
20. Hemphill JC 3rd, White DB. Clinical nihilism in neuroemergencies. Emerg Med Clin North Am. 2009;27(1):27–37, vii-viii. [PubMed: 19218017]
21. Park SY, Kuo P-Y, Barbarin A, et al. Identifying Challenges and Opportunities in Human-AI Collaboration in Healthcare. Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing; 2019; Austin, TX, USA.
22. Bansal G, Nushi B, Kamar E, Weld DS, Lasecki WS, Horvitz E. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. Proceedings of the AAAI Conference on Artificial Intelligence. 2019;33(01):2429–2437.
23. De-Arteaga M, Fogliato R, Chouldechova A. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; 2020; Honolulu, HI, USA.
24. Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General. 2015;144(1):114–126. [PubMed: 25401381]
25. Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012;19(1):121–127. [PubMed: 21685142]
26. Lebovitz S, Lifshitz-Assaf H, Levina N. To Incorporate or Not to Incorporate AI for Critical Judgments: The Importance of Ambiguity in Professionals' Judgment Process. NYU Stern School of Business. 2020.

27. Steinberg A, Grayek E, Arnold RM, et al. Physicians' cognitive approach to prognostication after cardiac arrest. Resuscitation. 2022.
28. Reynolds JC, Grunau BE, Rittenberger JC, Sawyer KN, Kurz MC, Callaway CW. Association Between Duration of Resuscitation and Favorable Outcome After Out-of-Hospital Cardiac Arrest: Implications for Prolonging or Terminating Resuscitation. Circulation. 2016;134(25):2084–2094. [PubMed: 27760796]
29. Drennan IR, Case E, Verbeek PR, et al. A comparison of the universal TOR Guideline to the absence of prehospital ROSC and duration of resuscitation in predicting futility from out-of-hospital cardiac arrest. Resuscitation. 2017;111:96–102. [PubMed: 27923115]
30. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–453. [PubMed: 31649194]
31. Campbell BC, Mitchell PJ, Kleinig TJ, et al. Endovascular therapy for ischemic stroke with perfusion-imaging selection. N Engl J Med. 2015;372(11):1009–1018. [PubMed: 25671797]
32. Hacke W, Kaste M, Bluhmki E, et al. Thrombolysis with alteplase 3 to 4.5 hours after acute ischemic stroke. N Engl J Med. 2008;359(13):1317–1329. [PubMed: 18815396]
33. Nogueira RG, Jadhav AP, Haussen DC, et al. Thrombectomy 6 to 24 Hours after Stroke with a Mismatch between Deficit and Infarct. N Engl J Med. 2018;378(1):11–21. [PubMed: 29129157]
34. Nieuwkamp DJ, Setz LE, Algra A, Linn FH, de Rooij NK, Rinkel GJ. Changes in case fatality of aneurysmal subarachnoid haemorrhage over time, according to age, sex, and region: a meta-analysis. Lancet Neurol. 2009;8(7):635–642. [PubMed: 19501022]
35. Gerke S, Babic B, Evgeniou T, Cohen IG. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. npj Digital Medicine. 2020;3(1):53. [PubMed: 32285013]
36. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. npj Digital Medicine. 2021;4(1):4. [PubMed: 33402680]
37. Raghu M, Blumer K, Sayres R, et al. Direct Uncertainty Prediction for Medical Second Opinions. Proceedings of the 36th International Conference on Machine Learning; 2019; Proceedings of Machine Learning Research.
38. Perdomo J, Zrnic T, Mendler-Dünner C, Hardt M. Performative Prediction. Proceedings of the 37th International Conference on Machine Learning; 2020; Proceedings of Machine Learning Research.
39. Coston A, Kennedy EH, Chouldechova A. Counterfactual Predictions under Runtime Confounding. ArXiv. 2020;abs/2006.16916.
40. Schulam P, Saria S. Reliable decision support using counterfactual models. Advances in neural information processing systems. 2017;2017-December:1698–1709.
41. Gigerenzer G, Gaissmaier W, Kurz-Milcke E, Schwartz LM, Woloshin S. Helping Doctors and Patients Make Sense of Health Statistics. Psychological Science in the Public Interest. 2007;8(2):53–96. [PubMed: 26161749]
42. Geocadin RG, Callaway CW, Fink EL, et al. Standards for Studies of Neurological Prognostication in Comatose Survivors of Cardiac Arrest: A Scientific Statement From the American Heart Association. Circulation. 2019;140(9):e517–e542. [PubMed: 31291775]
43. Scarpino M, Lolli F, Lanzo G, et al. Neurophysiology and neuroimaging accurately predict poor neurological outcome within 24 hours after cardiac arrest: The ProNeCA prospective multicentre prognostication study. Resuscitation. 2019;143:115–123. [PubMed: 31400398]
44. Rudin C Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019;1(5):206–215.
45. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. Proceedings of the 4th Machine Learning for Healthcare Conference; 2019; Proceedings of Machine Learning Research.
46. Kaur H, Nori H, Jenkins S, Caruana R, Wallach H, Vaughan JW. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery; 2020:1–14.

47. Lakkaraju H, Bastani O. “How do I fool you?”: Manipulating User Trust via Misleading Black Box Explanations. 2019:arXiv:1911.06473. https://ui.adsabs.harvard.edu/abs/2019arXiv191106473L. Accessed November 01, 2019.

48. Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Vaughan JWW, Wallach H. Manipulating and Measuring Model Interpretability. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery; 2021:Article 237.

49. De-Arteaga M, Dubrawski A, Chouldechova A. Learning under selective labels in the presence of expert consistency. ArXiv. 2018;abs/1807.00905.

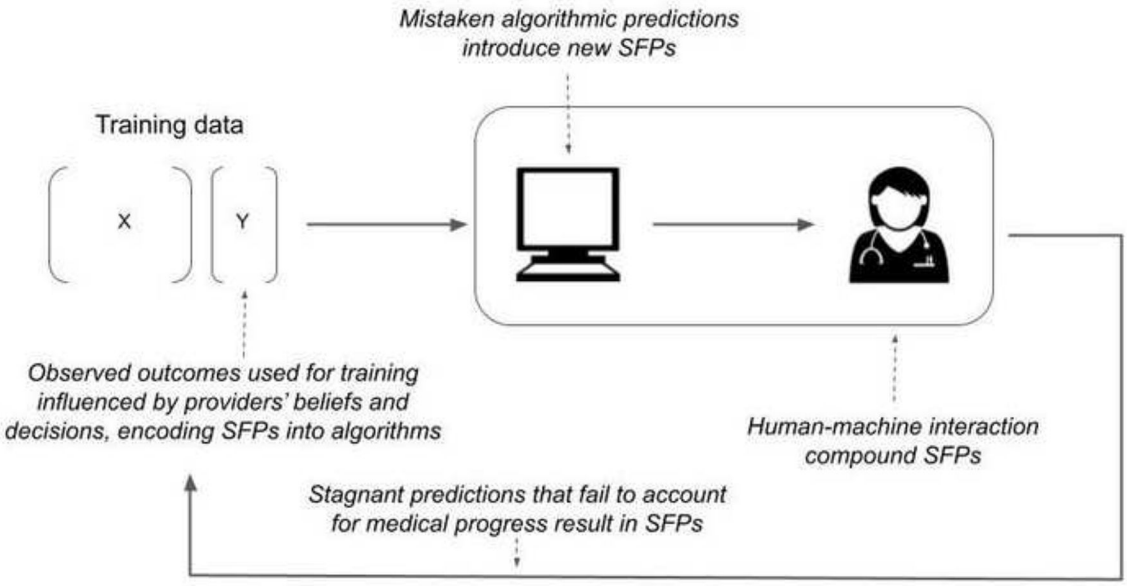

Figure 1:
Conceptual diagram of how self-fulfilling prophecies can be created, perpetuated, and amplified through machine learning and algorithmic prediction.

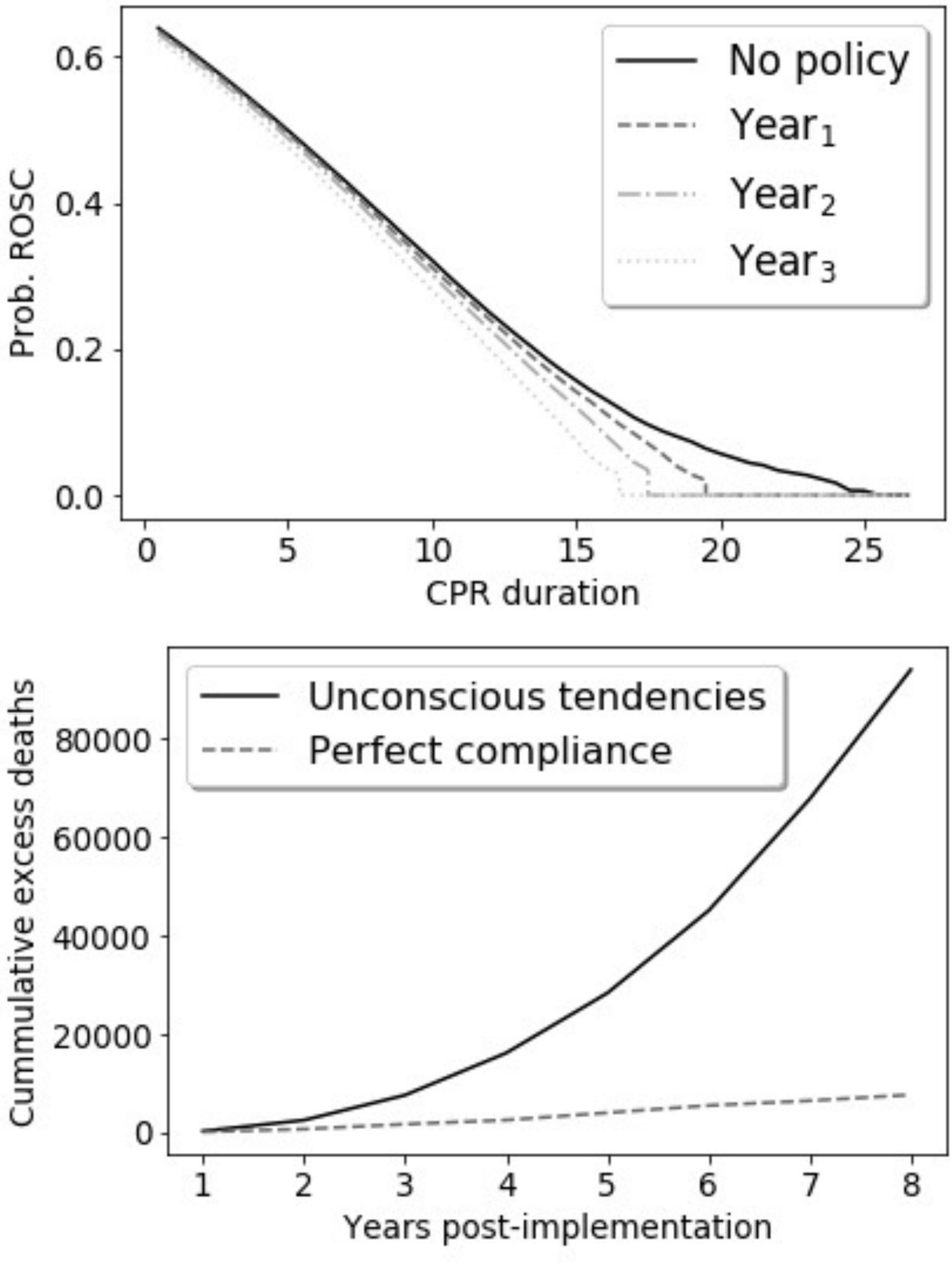

Figure 2:

Results of data simulations for treatment guidelines that recommend TOR when it is estimated that probability of ROSC is below 5%. We assume guidelines are updated yearly based on observational data. Full simulation setup and results are available in

Supplemental Appendix 1. Panel A: Excess deaths, compared to deaths in the absence of a policy recommending TOR. Panel B: Observed probability of ROSC under guidelines recommending TOR and presence of tendencies (e.g., nihilism) impacted by the guidelines.