Draft Paper, 31 July 2023

Making a Murderer –

How risk assessment tools may produce rather than predict criminal behavior

Donal Khosrowi and Philippe van Basshuysen

Abstract

Risk assessment tools, such as COMPAS, may not only predict defendants' risk of recidivism, but may themselves causally affect recidivism outcomes. We argue that such "performative" effects can yield severe harms both to individuals and to society at large, which raise epistemic-ethical responsibilities on the part of developers and users of risk assessment tools. To meet these responsibilities, we present a novel desideratum on algorithmic tools, called *explainability-in-context*, which requires clarifying how these tools causally interact with the social, technological, and institutional environments they are embedded in. Risk assessment practices are thus subject to high epistemic standards, which haven't been sufficiently appreciated to date. Explainability-in-context, we contend, is a crucial goal to pursue in addressing the ethical challenges surrounding risk assessment tools.

Keywords: performativity; risk assessment; recidivism; algorithmic fairness; explainability

1 Introduction

Algorithmic risk assessment tools are increasingly used in criminal justice systems around the world to help judicial actors evaluate defendants' risk of reoffending in the future and to inform sentencing decisions. There has been extensive debate in recent years about whether these tools exhibit problematic kinds of biases, including racial bias (e.g. Angwin et al. 2016; Biddle 2020; Dieterich et al. 2016; Flores et al. 2016; Freeman 2016; Hedden 2021; Hellman 2020; Washington 2019). In this paper, we highlight a related but distinct problem surrounding the use of risk assessment tools: *performativity*.

A risk assessment tool (RAT) is performative when its predictions about a defendant's future behaviors causally affect those behaviors. Performativity can arise in different forms, such as when a prediction is self-defeating or self-fulfilling (Khosrowi 2023). We focus on a particularly problematic version, where 1) a defendant's risk to reoffend is predicted as high, 2) the defendant is subsequently incarcerated, 3) the defendant engages in future criminal behaviors in line with the prediction, but 4) incarceration itself is the primary cause that induces these behaviors. In such a case, the RAT correctly classifies the defendant as high-risk, but it does so for the wrong reasons because the prediction is self-fulfilling in virtue of the criminogenic effects that incarceration can have (e.g. Lambie and Randell 2013; Stevenson 2017).

While we are not the first to note the potentially self-fulfilling nature of risk assessments (e.g. Morgan et al.; Hamilton 2015a; Sidhu 2015), we argue that this type of performativity creates previously underappreciated epistemic-ethical challenges. First, because it imposes significant injustices on defendants and creates harms for society at large. Second, because such injustices and harms are difficult to recognize empirically: the predictive performance of risk assessment tools is routinely validated against observational data, but this approach fails to trace *why* there is agreement between predictions and actual courses of events.

The performativity of RATs hence points to serious deficiencies in the methodology used to evaluate these tools. To address these problems, we argue that RATs need to be complemented with sincere efforts aimed at *explainability*. Explainable artificial intelligence (XAI) seeks to help stakeholders understand how and why algorithmic systems come to make certain predictions, classifications or recommendations and continues to attract significant

attention from AI researchers, ethicists, civil rights scholars and many others (e.g. Burrell 2016; Fleisher 2021; Selbst and Barocas 2018; Winter, Hollman, and Manheim 2023). Here, we propose a wider explainability desideratum we call *explainability-in-context* (EIC). In a nutshell, EIC aims to elucidate how an AI system interacts with the wider social, technological, and institutional landscape it is embedded in. In particular, we advocate that validating RATs requires taking a causal perspective that aims to understand how these tools causally interact with a target to co-shape the very outcomes they serve to predict. In the case of RATs, an ideal version of this strategy would seek to distinguish between prediction-dependent and prediction-independent causes of criminal behavior, and only allow the latter to bear on predictions regarding a defendant's future behavior.

More broadly, an EIC perspective grounds the claim that performativity yields previously underappreciated epistemic-ethical responsibilities on the part of developers and users of RATs: these tools aren't merely *epistemic* tools that provide one-way access to decision-relevant features of the world, but are *performative* tools that have the capacity to change the world, for better or worse. Developers and users must hence 1) epistemically, ensure that RATs only predict behaviors that are prediction-independent, and 2) morally, be accountable for any harms induced through departures from this constraint.[1] While additional work is needed to study the challenges raised by performativity, to develop ways to alert institutional actors to its significance and the epistemic-ethical responsibilities that arise from it, and to

[1] Note that our argument doesn't amount to a principled objection to the use of RATs in the criminal justice system. Rather, we believe that in the absence of these tools, possibly harmful performative effects may materialize through subjective risk assessments "inside decision makers' heads". For instance, a judge might form an opinion about a defendant's recidivism risk being high, which may lead her to impose a severe sentence, thus triggering a chain of events that will make her risk assessment come true. But, while performative effects following subjective assessments may incorporate various kinds of biases that will be hard to scrutinize and to properly account for, RATs, in contrast, promise to make it in principle possible to understand their own causal effects on decisions and thus, to rule out harmful effects. To deliver on this promise, however, RATs should be made explainable-in-context, as we argue below.

design better strategies to mitigate harmful forms of performativity, EIC marks a first important step in guiding such efforts.

The discussion is organized as follows. *Section 2* elaborates on the problem of performative predictions in the context of RATs in the criminal justice system. We identify the two main causal pathways that enable harmful performativity: 1) RATs can causally influence the type and severity of sentences, and 2) sentencing decisions can in turn causally affect downstream criminal behavior. Against this background, we highlight how harmful performative effects may arise and go unnoticed when validating the predictive performance of RATs from observational data. *Section 3* outlines EIC as a general desideratum and elaborates on what concrete demands it imposes for dealing with performative predictions responsibly. *Section 4* articulates the epistemic-ethical responsibilities arising on the part of developers and users to promote EIC for RATs. *Section 5* concludes.

2 Can risk assessments become self-fulfilling prophecies?

Consider the following example:

> Terrell, an 18-year old male from a disadvantaged neighborhood, is arrested for the possession of methamphetamine with intent to distribute. Even though he has no prior criminal records, Terrell's risk of recidivism – including violent recidivism – is assessed as high by a RAT. Based on the assessment, the judge rules out probation and passes the maximum sentence of five years in state prison. During his first year in prison, a cell mate blackmails Terrell and extorts money from him. Desperate to escape his predicament, Terrell joins a gang that promises protection in exchange for

executing jobs for the gang. Within six months upon his release, Terrell fatally shoots a man who is said to be a debtor of the gang.

In our vignette, the risk assessment becomes a self-fulfilling prophecy, because it triggers a chain of events that leads to Terrell shooting the man. Had the judge not based the verdict on the high estimated recidivism risk, Terrell may have been released on probation. He wouldn't have been blackmailed, wouldn't have joined the gang, and subsequently, he wouldn't have shot the victim. The example is stylized in order to enable us to clearly trace the causal chain of events, from the high risk assessment to the harsh sentence, and from Terrell's stint in prison to the shooting. In any real case, however, identifying similar causal links may be difficult and contestable. So, may real-world risk assessments actually be self-fulfilling in ways similar to our fictitious case? Rather than seeking to identify concrete cases where this might have happened, our aim here is to provide a proof-of-concept by identifying two causal pathways which, taken together, establish a possible way in which risk assessments can be self-fulfilling: first, risk assessments may affect decisions regarding probation and the severity of sentences; and second, time spent in prison may have criminogenic effects. Let's consider these causal pathways in turn.

2.1 Do risk assessments affect sentences?

Whether risk assessments may be used to inform sentencing is a controversial issue in ethics, law, and public policy. When determining whether someone will be sent to prison, or for how long, the permissibility of drawing on risk assessments will depend on the theory of sentencing that the sentencing decision is based on. According to some theories, sentences ought to be exclusively determined by what should be regarded as the fair retribution for a crime, which rules out the use of forward-looking risk assessments in sentencing decisions

(see Monahan and Skeem 2014; 2016). Other theories, in contrast, base sentencing on the consequences for the defendant and society. These theories thus allow for a more forward-looking sentencing including the use of risk assessment, as do hybrid theories that combine retributivist and forward-looking considerations in sentencing (see Monahan and Skeem 2014; 2016; Hamilton 2015b). But, even when we find forward-looking punishment permissible, as many influential theories of sentencing (e.g. Morris's 1974) in fact do, this doesn't imply that the use of concrete tools, such as COMPAS, is permissible for sentencing as, for instance, these tools might not be designed or fit for such purposes (cf. State vs. Loomis 2016, 100).

Notwithstanding prevailing disagreement, actual practice commonly includes information from RATs in pre-sentence investigation reports (PSIs) that are provided to sentencing courts in the US (Casey et al. 2011), and PSIs appear to be widely used in helping to arrive at sentencing decisions. This is fully licensed in many US states; as Monahan and Skeem (2014) point out, a number of states have explicitly "[...] incorporated risk assessment into sentencing guidelines as one factor that judges *may* consider in determining the appropriate sentence within the limits established by law" (2014, 495, emphasis original).

A much-discussed case where information from RATs was used is that of Eric L. Loomis.[2] Loomis was accused of being the driver in a drive-by shooting. His PSI included COMPAS risk scores, according to which Loomis presented a high risk of recidivism, including of violent recidivism. Even though it was emphasized in the PSI that these scores shouldn't be used for determining probation or the severity of the sentence, it appears that the responsible circuit court, following an argument by the state attorney, did both. First, the court made reference to the COMPAS scores when ruling out probation:

[2] See State v. Loomis 2016; also Freeman 2016; Washington 2019.

> You're identified, through the COMPAS assessment, as an individual who is at high risk to the community. In terms of weighing the various factors, I'm ruling out probation because of the seriousness of the crime *and because* your history, your history on supervision, and the risk assessment tools that have been utilized, suggest that you're extremely high risk to re-offend. (State v. Loomis 2016, 19, emphasis added)

Based on the court's language, we take this case to indicate that information from COMPAS was at least among the factors that causally contributed to ruling out probation for Loomis.

Whether the court also used the scores for determining the length of the sentence is less clear. After receiving the maximum penalty on two counts (six years of imprisonment and five years of extended supervision), Loomis filed a motion soliciting a new hearing, arguing that the court's consideration of the COMPAS risk scores at sentencing violated his due process rights (State v. Loomis 2016, 23). Perhaps surprisingly, the court denied the motion not by arguing, against Loomis, that its use of the risk scores at sentencing didn't violate his due process rights. Rather, the court denied that the risk scores had contributed to determining Loomis's sentence in the first place, explaining it had only "used the COMPAS risk assessment to *corroborate* its findings and that it *would have imposed the same sentence regardless of whether it considered the COMPAS risk scores*" (ibid., 28, emphasis added). The court's reasoning here is unconvincing, however. To rule out that the scores contributed to determining the sentence, the relevant counterfactual to consider is not whether the court would have imposed the identical sentence regardless of whether it considered the COMPAS scores, as the court contended, but rather whether it would have imposed the identical sentence *had the scores been different*. Since the court remained silent about these counterfactuals, its reasoning fell short of responding to Loomis's due process challenges.

Furthermore, the Wisconsin Supreme Court indirectly confirmed the circuit court's use of COMPAS at sentencing, while however denying that this had violated Loomis's due process rights: "Ultimately, we disagree with Loomis because consideration of a COMPAS risk assessment at sentencing along with other supporting factors is helpful in providing the sentencing court with as much information as possible in order to arrive at an individualized sentence" (State v. Loomis 2016, 765). The Supreme Court's ruling thus licensed the use of COMPAS at sentencing, only circumscribing its use by requiring that "[a] COMPAS risk assessment is only one of many factors that may be considered and weighed at sentencing" (State v. Loomis 2016, 769). Summing up, the circuit court, in this case, made use of COMPAS scores in ruling out probation and, arguably, in determining the severity of the sentence; and the Wisconsin Supreme Court licensed this use.

Moving beyond State vs. Loomis as an example, risk assessments are routinely included in PSIs and are frequently referenced when determining probation or the severity of sentences (e.g. Malenchik v. State 2010), so it is plausible to think that RAT outputs are frequently involved in determining sentence type and severity/duration.

2.2 Does incarceration have criminogenic effects?

Let us turn to the second causal pathway, which concerns the effects of incarceration on recidivism. The criminogenic effects of incarceration have been investigated by researchers for decades and there is now substantial empirical and theoretical literature that documents and analyses such effects. For instance, a host of empirical studies have made progress on clarifying the criminogenic effects of imprisonment by estimating differences in recidivism between otherwise similar defendants with respect to whether they were incarcerated, their

sentence was suspended or they were sentenced to community services (standalone or as probation) (Cid 2009; Spohn and Holleran 2002; Vieraitis et al. 2007; see Stemen 2017 for an overview). Relatedly, other studies find similar effects for shorter pre-trial detention periods (Lowenkamp et al. 2013; Gupta et al. 2916; Heaton et al. 2017), suggesting that even short durations of incarceration can yield recognizable differences in defendants' future propensity to (re-)offend.

These empirical efforts are complemented by a wide range of attempts to explain the relationship between imprisonment and (future) criminal behavior. These range from atheoretical, common sense causal narratives to theory-driven attempts to explain how different factors bear on criminal behavior and how imprisonment, in turn, can intervene with these in negative ways. Broadly, imprisonment is often argued to be conducive to recidivism by disrupting family ties and offenders' social networks; worsening offenders' mental health, especially when imprisonment does not involve rehabilitative efforts; and by making it more difficult for offenders to secure housing and employment upon release. Moreover, especially for young offenders, imprisonment is often believed to cement a lack of education and to deprive offenders of opportunities to build social capital relative to their peers. Additional worries focus on the effects of being exposed to other inmates with high criminal potential, which may draw susceptible offenders towards criminal behaviors in the future. All of these factors, plausibly, can make it difficult for offenders to re-enter society in a way that protects them from being drawn towards criminal behaviors, such as in our Terrell vignette. Some of the concerns outlined here have been put on theoretical grounding, too. For instance, *labeling theory* (Paternoster and Iovanni 1989) seeks to describe the ways in which ex-prisoners' being labeled as such can drastically change not just formal but also informal interactions with society on re-entry, worsening offenders' prospects at successful re-integration in ways

that are difficult to detect, e.g. through subtle stereotyping and associated prejudice. While these attempts at explaining empirically diagnosed relationships between incarceration and criminal behavior seem largely compelling, we do not commit here to any of the more specific mechanisms offered. Our arguments remain effective even if the true causal relationships by which incarceration influences criminal behavior remain unknown.

In sum, there are both empirical and theoretical reasons to believe that incarceration can causally affect individuals' propensity to reoffend in the future. Because of this, using RATs to inform pre-trial judicial decisions and sentencing can sometimes yield self-fulfilling performative effects. We hasten to add that we are not the first to emphasize this as a problem: the performative potentials of risk assessment have been recognized in the existing literature, both in the general context of judicial decision-making as well as in the specific context of using RATs (Hamilton 2015; Barabas et al. 2018). Our arguments differ from existing contributions, however, in that we specifically focus on the epistemic-ethical problems that arise when RATs are presented and used as putatively predictive tools. On our account, RATs' functioning can significantly extend beyond this role and are better understood as performative instruments that do not only issue risk assessments but also have the capacity to actively shape the world, including in epistemic-ethically problematic ways. This in turn yields underappreciated epistemic-ethical obligations on the part of producers and users of RATs, as we will argue below.

2.3 Linking the pathways

We are now in a better position to understand the problems performativity can pose for the responsible use of RATs. The first causal pathway captures how RATs can causally

contribute to whether defendants are incarcerated and for how long. The second causal pathway, in turn, makes clear that risk assessments can yield self-fulfilling performative effects, *inducing* rather than just predicting future criminal behavior. Both pathways are *necessary* for harmful performative effects, but not yet sufficient.

To better understand how such effects can come about, let us consider some stylized cases which illustrate how a RAT implemented in a criminal justice system may come to induce harmful performative effects. Before RATs can be used in judicial decision making, they must be tested and validated to show that they can provide accurate assessments. Without such validation, decision-makers would likely not use RATs at all. But validating the predictive performance of RATs against available observational data is susceptible to turning features of defendants that are otherwise causally and probabilistically unrelated to recidivism into good predictors of it, though only in virtue of self-fulfilling performative effects. Let us consider two cases in turn to illustrate how this might happen.

Assume that judicial decisions about pre-trial detention, parole and imprisonment in some past period (say, the 1950s-1980s) were widely based on a range of individual characteristics decision-makers believed to be relevant to defendants' recidivism propensity. Let us assume that at least some of the variables were neither causally nor statistically relevant for recidivism but were still frequently included to make risk assessments. Let us take one variable, G, which tracks a defendant's gender and assume that some fraction of decision makers believed that males were more likely to reoffend than females, all else equal. Let us also imagine that this assumption was wrong, perhaps induced by confirmation bias, cherry-picking of notable cases, and so on. So, gender may have been a tie-breaker between defendants being incarcerated or not in a substantial number of cases, but wrongly so. Why is this a problem? First, if gender does not have incarceration-independent statistical or causal relationships to recidivism, then it is simply wrong, epistemically (as well as morally and

legally), for judicial decision-makers to have used it in decision-making. But, more pertinently for our present arguments concerning RATs, using gender in this way can also *establish* a statistical relationship between gender and recidivism that becomes important when developing and validating RATs later on. Let's fast forward in our stylized mini-world to the 2000s. If a developer draws on observational data of defendant characteristics and post-trial criminal history, they will see that gender and recidivism are correlated, perhaps even when controlling for other factors. And this is not just a statistical fluke. If gender played an important role in prior sentencing decisions, and incarceration has criminogenic effects, then gender is indeed a good predictor of recidivism because it is now causally relevant for recidivism. But, problematically, it is the wrong variable to focus on. In an ideal setting, with varied and large enough datasets, 'controlling for' imprisonment status, or sentence type could, at least in principle, shed some light on whether the correlation between gender and recidivism is due to imprisonment(-type) or not. But unfortunately, no risk assessment score or tool we are aware of has been developed to take this into account. Pertinently, the development and validation of Equivant's controversial COMPAS tool was "strongly influenced" (Equivant 2019, 13) by the OGRS (offender group reconviction scale) score offered by Copas and Marshall (1998). While Copas and Marshall explicitly consider criminogenic effects of incarceration, they note that their approach "[...] does not condition on the sentence given and so if a sentence has an effect on reconviction then this effect is not taken into account" (ibid, 170). Equivant (2019) do not comment at all on whether the estimation strategy underlying COMPAS improves on Copas and Marshall's approach, so there are reasons to worry that it does not.

Importantly, the case described here is one where a RAT itself is faultless insofar as it does not itself induce the relationship between gender and recidivism. Its malfunctioning is only parasitic on earlier, human decision-making, which baked this relationship into the

observational data a RAT is calibrated on and tested against. But we can easily imagine more severe cases where a RAT itself induces such statistical relationships. Social science methodologists have known for more than a century that spurious correlations are a threat to sound predictive (and causal) inference, so all it takes is a variable Z that a developer considers a candidate for predicting recidivism, and once Z is included in a RAT that is deployed, this may cement or intensify Z's predictive relevance by virtue of performative effects.

In sum, our concern is that even variables that are, in principle, causally and statistically unrelated to recidivism can end up being statistically relevant for predicting recidivism if prior decision-making has considered them relevant for decisions about sentence type, and if sentence type, in particular imprisonment, can be causally relevant to increasing recidivism. In such a case, a RAT may *successfully* predict recidivism, but for the wrong reasons. Performativity is hence a crucial concern for validating the predictive accuracy of RATs. Existing methods for RAT validation do not, to our knowledge, consider performative effects, but unless they do, we may reasonably be suspicious about whether the use of RATs in the criminal justice system is warranted. Let us turn to explicating in a more systematic way what principled methodology could be useful in making progress. As we argue, 1) it is empirically challenging, but not impossible, to explore whether RATs are performative, 2) developers and users must meet these challenges by pursuing the ideal of explainability-in-context (EIC), and 3) if they don't, they may violate important epistemic-ethical obligations that undermine the legitimacy of using RATs.

3 Explainability-in-Context

The central problem posed by self-fulfilling risk assessments is now clearly in view. They can induce, cement, and intensify statistical and causal relationships between individuals' characteristics and recidivism, leading them to make predictions about recidivism that may be accurate, but for the wrong reasons. If unnoticed and unaddressed, the performativity of RATs can hence impose substantial but possibly avoidable harm on defendants and society at large. So, what should we do to address this challenge?

Here, we propose a novel epistemic-ethical desideratum, which we call *explainability-in-context* (EIC). In a nutshell, EIC involves exploring, understanding and explaining how algorithmic systems interact with the broader environment they are embedded in. EIC draws on two established movements. One is the socio-technical systems approach popular in science and technology studies, which aims to understand how technologies interact with the social and institutional settings they are implemented in and seeks to identify ways to co-ordinate their interaction and integration in beneficial ways (Pitt, Schaumeier, and Artikis 2012; Chopra and Singh 2016; Chopra and Singh 2018). The second is the recent cross-disciplinary movement calling for explainability of artificial intelligence (AI) systems (Burrell 2016; in the context of the judicial system see Selbst and Barocas 2018; Atkinson, Bench-Capon, and Bollegala 2020; Deeks 2020; Winter, Hollman, and Manheim 2023). These systems are often argued to be inscrutable black boxes that do not readily reveal information about their internal functioning, making it difficult for stakeholders to assess whether they function in acceptable or problematic ways. In the case of RATs, for instance, critics stress it often remains unknown what information is used to make an assessment and how different characteristics are weighted, as these aspects are veiled as trade secrets relating to proprietary technologies (Freeman 2016; Washington 2019; Biddle 2022). Some take this inscrutability to be threatening defendants' right to due process, as without knowing what

information is used and in what ways, it is not possible for defendants to challenge risk assessments (Freeman 2016; Washington 2019; Biddle 2022). Extending the scope of explainability, and in the spirit of the sociotechnical systems approach, EIC insists that we must be able to understand how RATs like COMPAS are likely to *causally affect* defendants, not just which variables inform an assessment and how they are weighted. EIC then amounts to understanding how systems like COMPAS interact with the criminal justice system, starting from what data are used to estimate risk models and extending to how risk scores influence decision-making and how the use of these scores causally affects the outcomes to be predicted.

How, then, can we make progress towards EIC for RATs? The main problem we see with RATs in a performative world is that existing methodology to develop and validate RATs does not, so far, distinguish between prediction-dependent and prediction-independent effects on the outcomes to be predicted, e.g. between a defendant who would have recidivated regardless of sentence-type and a defendant who would recidivate if imprisoned but wouldn't if sentenced to probation. This distinction is crucial for understanding whether RATs that predict accurately do so for the wrong reasons, and induce rather than merely predict outcomes. Ideally, RATs should be able to make assessments that discern between factors that bear on recidivism only through sentence-type and those that influence recidivism regardless of sentence-type. For instance, membership in a white supremacist terrorist organization might be a good predictor of recidivism if it encodes that individuals are likely to reintegrate into their gang upon release and continue to engage in joint criminal endeavors. In this case, it doesn't matter for someone's recidivism risk whether they receive a probationary or a custodial sentence for their racist attack on a middle eastern store clerk – their propensity to reoffend in similar ways will remain the same. Contrast this with a case

where, let us imagine, defendants with blue hair have been especially likely to be incarcerated and incarceration is criminogenic. In such a case, blue hair will correlate with recidivism and hence be valuable for predicting it, but is unrelated to recidivism other than *through* imprisonment. What distinguishes both cases is that when 'controlling for' or 'conditioning on' imprisonment (or sentence type more generally), blue hair would be unrelated to recidivism, whereas membership in a white supremacist terrorist organization would not be. The latter may remain predictively relevant even if we control for the effects that imprisonment may have on recidivism, whereas the former's statistical relationship to recidivism would be filtered out. A useful way of teasing apart prediction-dependent and prediction-independent effects is to consider some causal graphs and their associated probabilistic dependencies and independencies. *Figure 1* captures causal diagrams for the two cases.

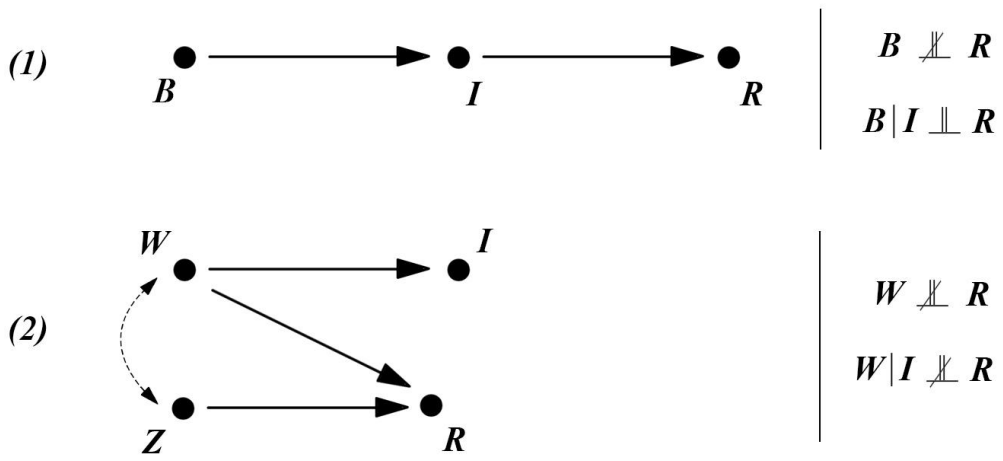

Figure 1. Causal diagrams for prediction-dependent and prediction-independent effects

Here, blue hair (B) is causally relevant for recidivism (R), but only through imprisonment (I). By contrast, being a member of a white supremacist organization (W) is both directly

causally relevant for R, e.g. because it encodes attitudes that directly cause violent attacks against minority individuals and because it is correlated (dashed arc) with crime-conducive causal factors, e.g. gun-ownership, being in contact with other individuals prone to commit violent offenses (Z), which are in turn directly relevant to R[3]. Importantly, both B and W are probabilistically dependent on I unconditionally, so they are indistinguishable through this narrow lens. However, once we condition on I, B is rendered independent, whereas W remains dependent, suggesting they play different causal roles. This difference helps us recognize that using blue hair to predict recidivism may problematically involve prediction-dependent effects, whereas white supremacist status does not raise this problem. Understanding how a RAT latches onto different kinds of correlates and causes of recidivism is hence crucial for understanding whether its predictive performance rests on problematic reasons.

Our illustrations here are naturally limited - they do not constitute a definitive method for teasing apart prediction-dependent and independent effects. It is widely understood that casually interrogating probabilistic dependencies is a crude way of understanding what goes on causally. We have taken this approach here simply for ease of illustration and note that more advanced methods are available to disambiguate causal relationships from observational data (e.g. Pearl 2009). Ultimately, such efforts should pursue a common objective. EIC demands that RATs are able to make (accurate) counterfactual predictions. For each case, they should be able to predict R for at least two scenarios: one telling us what an individual's outcome would be if they were imprisoned, and the other telling us what would happen if they were subjected to a different sentence type, e.g. probation, community service etc. If a

[3] Note that we assume that there is no I → R arrow for the white supremacist, encoding that they are criminally 'saturated', so imprisonment does not further increase their post-sentence recidivism probability.

decision-maker recognizes risk scores to be robust over the decision-space, they may feel more confident that a RAT tracks prediction-independent causes of R. However, if the scores change over the decision-space, this can alert them to the performative nature of the assessment. Likewise, defendants (and their legal representatives) should be able to interrogate risk scores on these aspects, and challenge decisions based on them if performative effects cannot be ruled out. As it stands, RATs do not have capabilities to make such counterfactual predictions, to our knowledge, and it seems that developing these is a first important step towards helping relevant stakeholders better understand how RATs function and if this proceeds in epistemically, morally and legally acceptable or problematic ways.

In sum, looking at performativity through a causal lens reveals a major deficiency of existing methods for developing and validating RATs. To tell whether a prediction may include prediction-dependent effects, we should look at counterfactuals, but unfortunately this is not done in practice. Importantly, if these predictions differ, this means that there are performative effects that we may wish to exclude in furnishing our prediction.

4 Epistemic-ethical duties of developers and users of RATs

EIC's emphasis on understanding how RATs causally interact with the criminal justice system raises duties on the part of developers and users of RATs, at various stages of the development and deployment of these tools.

First, developers should ensure RATs allow for conditional predictions, including the performative effects of issuing different kinds of sentences. In particular, RATs should

explicitly integrate the effects that a high risk assessment can have on individuals by means of harsher sentences. Performativity is a familiar problem encountered in other areas, too. For instance, in the machine learning literature, computer scientists have proposed strategies to incorporate the causal effects of a prediction into the prediction itself (Perdomo et al. 2021; see also Hardt et al. 2022). The aim here is mainly to deal with self-effacing predictions, where a models' performativity undermines its predictive accuracy, e.g. because agents respond to a prediction. This is different from the cases we highlight, where performativity is self-fulfilling. Here, the aim is not to calibrate a RAT to incorporate its own, crime-inducing effects; the problem is rather that these might already be part of RAT prediction, and the aim is to tease prediction-dependent and prediction-independent effects apart and to issue *different* conditional predictions for different scenarios in a way that reflects performative effects and makes them accessible to decision-makers. So while not quite yet in tune with the needs arising to manage performativity of RATs, we believe that future work drawing on recent approaches offered by computer scientists can make important steps towards EIC.
In sum, since it seems in principle possible to provide conditional predictions comparing the effects of different kinds of sentences, we contend, providers of RATs have a duty to follow this strategy (and to develop techniques that help with doing so), and, conversely, should be held accountable for any harms induced through departures from this desideratum.

Second, users of RATs have the duty to use these tools in responsible ways. In particular, pursuing EIC enables them to understand whether using the system might have undesirable or unintended effects, which they ought to reflect upon and account for in their decision making. There will mainly be two relevant user groups. The first consists of judges and jury members. Faced with conditional predictions, these users ought to reflect on the conditional predictions provided through EIC to better understand the causal pathways their sentencing decisions

may influence, possibly in detrimental ways. In particular, by providing predictions about how their sentencing decisions might influence a defendant's rehabilitation and recidivism outcomes, EIC may ultimately prevent them from unjustly incarcerating a defendant. Secondly, for defendants and their attorneys, EIC promises to help provide understanding of how judgments and predictions came about and may help challenge sentences when these are based on epistemically or ethically inappropriate grounding. For instance, with the help of conditional predictions, it might convincingly be argued that a prediction may be conducive to becoming self-fulfilling when contributing to harsh sentencing decisions, which may in turn constitute a violation of due process. Here, EIC complements existing efforts to increase the transparency of algorithmic decision-making systems like COMPAS, such as Rudin et al. (2020), and help stakeholders interrogate these systems for relevant properties, such as fairness. EIC adds to this by widening the scope to not only consider how systems like COMPAS work internally, but also how they interact with the environments they are deployed in.

At this point, several objections may be raised concerning the epistemic duties that EIC imposes on providers and users of RATs. First, we have focused here on how EIC may help prevent harmful cases where high risk assessments unjustly lead to higher rates of imprisonment and this goes unnoticed as the assessments constitute self-fulfilling prophecies. But other, beneficial cases seem possible, too, e.g., where, without risk assessment, a defendant might be incarcerated, subject to the criminogenic effects of imprisonment, but a hypothetical low risk assessment would, performatively, prevent these effects from obtaining and thus lead to more favorable outcomes. So, given that RATs may also have beneficial performative effects, wouldn't EIC seem to prevent such benefits from obtaining? To clarify, EIC is not intended as a desideratum on how to *manage* the performativity of RATs: it does

not insist that performative effects need to be mitigated. It is rather an epistemic-ethical desideratum to govern how RATs should be developed and used, i.e. with a view towards exploring and understanding performative effects. By helping provide agents with epistemic access to such effects, harmful and beneficial, EIC does not make specific recommendations for whether performative effects should be mitigated. This, ultimately, must be the subject of a larger debate informed by substantive ethical, legal and social theory, including about the functions of the criminal justice system and its role in society.

A second, related concern is that EIC requires a very high epistemic bar that must be met before RATs can justifiably be implemented. Meeting this epistemic bar will raise costs and risks for companies developing these tools, and for the criminal justice system more generally. But, if these costs and risks are too high, this might prevent RATs from being developed and used in the first place, which means, however, foregoing their possible benefits, too. Surely, when evaluating and implementing these tools, we should consider both their risks and their benefits (see van Wijck 2013), but the epistemic duties implied by EIC may skew the calculation towards a very conservative use of RATs. In response, we do not argue that the high epistemic bar must be met before RATs may permissibly be implemented, or that we need a moratorium for RATs. As we have stressed before, risk assessments, with or without RATs, may have performative effects. But, while informal assessments may be performative in obscure ways and might encode subjective preferences and biases, RATs, when they are combined with an EIC approach, hold significant promise to provide a better understanding of such performative effects and to allow judicial actors to both mitigate their risks and reap their benefits. As RATs are increasingly used in the justice system, sincere efforts should be made towards achieving EIC.

5 Conclusion

We have argued that RATs should be viewed not as mere predictive but as *performative* tools that have the potential to influence offenders' future lives, for better or worse. In particular, high risk assessments may constitute self-fulfilling prophecies by way of harsher sentences and the subsequent criminogenic effects of imprisonment, which can constitute burdens placed both on offenders and society at large. This raises serious ethical concerns about the use of these tools. Our proposed solution, EIC, is to take a causal perspective, distinguishing prediction-dependent from prediction-independent causes of recidivism, and allowing only the latter to influence risk assessments. Developers and users of RATs have epistemic-ethical duties to prevent harmful performative effects from materializing, by working towards a full-fledged EIC-approach. Currently, these duties are being violated.

References

Angwin, Julia, Jeff Larson, Surya Mattu and Lauren Kirchner. 2016. "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks." *ProPublica*, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Atkinson, Katie, Trevor Bench-Capon, and Danushka Bollegala. 2020. "Explanation in AI and Law: Past, Present and Future," Artificial Intelligence 289:103387

Biddle, Justin. 2022. "On Predicting Recidivism: Epistemic Risk, Tradeoffs, and Values in Machine Learning." *Canadian Journal of Philosophy* 52(3):321-341.

Burrell, Jenna. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1). https://doi.org/10.1177/2053951715622512

Casey, Pamela M., Roger K. Warren, and Jennifer K. Elek. 2011. "Using Offender Risk and Needs Assessment Information at Sentencing: Guidance for Courts from a National Working Group." *National Center for State Courts*, https://www.ncsc.org/__data/assets/pdf_file/0019/25174/rna-guide-final.pdf

Chopra, Amit K., and Munindar P. Singh. 2016. "From Social Machines to Social Protocols: Software Engineering Foundations for Sociotechnical Systems," In Proceedings of the 25th International World Wide Web Conference. ACM, Montréal, 903–914.

Chopra, Amit K., and Munindar P. Singh. 2018. "Sociotechnical Systems and Ethics in the Large," In 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES'18), February 2–3, 2018, New Orleans, LA, USA. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3278721.3278740

Copas, John, and Peter Marshall. 1998. "The offender group reconviction scale: a statistical reconviction score for use by probation officers." *Journal of the Royal Statistical Society* Series C (Applied Statistics), 47(1):159-171.

Deeks, Ashley. 2020. "Secret Reason-Giving," *Yale Law Journal* 129(3):612-689

Freeman, Katherine. 2016. "Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in State v. Loomis." *North Carolina Journal of Law and Technology* 18:75-106.

Hamilton, Melissa. 2015a. "Back to the Future: The Influence of Criminal History on Risk Assessments." *Berkeley Journal of Criminal Law* 20(1):75-133.

Hamilton, Melissa. 2015b. “Risk-Needs Assessment: Constitutional and Ethical Challenges.” *American Criminal Law Review* 231, U of Houston Law Center No. 2014-W-2, Available at SSRN: https://ssrn.com/abstract=2506397 or http://dx.doi.org/10.2139/ssrn.2506397

Hardt, Moritz, Meena Jagadeesan, and Celestine Mendler-Dünner. 2022. “Performative Power.” arXiv:2203.17232v1 [cs.LG] 31 Mar 2022

Khosrowi, Donal. 2023. “Managing Performative Models.” *Philosophy of the Social Sciences* *0*(0). https://doi.org/10.1177/00483931231172455

Monahan, John, and Jennifer L. Skeem. 2014. “Risk redux: the resurgence of risk assessment in criminal sanctioning.” *Federal Sentencing Reporter* 26:158-166.

Monahan, John, and Jennifer L. Skeem. 2016. “Risk Assessment in Criminal Sentencing.” *Annual Review of Clinical Psychology* 12:489-513.

Morgan, Frank, Neil Morgan, and Irene Morgan. Risk Assessment in sentencing and corrections. A report to the Criminology Research Council for research project 22/95-96. https://www.aic.gov.au/crg/reports/crg-2295-6

Pearl, Judea. 2009. *Causality: Models, Reasoning, and Inference*. 2nd edition. New York: Cambridge University Press.

Perdomo, Juan C., Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2021. “Performative Prediction.” arXiv:2002.06673v4 [cs.LG] 26 Feb 2021

Pitt, Jeremy, Julia Schaumeier, and Alexander Artikis. 2012. “Axiomatization of socio-economic principles for self-organizing institutions: Concepts, experiments and challenges,” ACM Transactions on Autonomous and Adaptive Systems (TAAS) 7, 4 (Dec. 2012), 39:1–39:39.

Rudin, Cynthia, Caroline Wang, and Beau Coker. 2020. The Age of Secrecy and Unfairness in Recidivism Prediction. Harvard Data Science Review, 2(1). https://doi.org/10.1162/99608f92.6ed64b30

Selbst, Andrew D., and Solon Barocas. 2018. "The Intuitive Appeal of Explainable Machines," Fordham Law Review, vol. 87, no. 3, pp. 1085-1139

Sidhu, Dawinder S. 2015. "Moneyball Sentencing." *Boston College Law Review* 56:671-731.

STATE v. LOOMIS. 881 N.W.2d 749. https://www.leagle.com/decision/inwico20160713i48

van Wijck, Peter. 2013. "The economics of pre-crime interventions." *European Journal of Law and Economics* 35:441-458.

Washington, Anne. (2019). "How to Argue with an Algorithm: Lessons from the COMPAS ProPublica Debate." *The Colorado Technology Law Journal*, 17(1) http://ctlj.colorado.edu

Winter, Christoph, Nicholas Hollman, and David Manheim. 2023. "Value alignment for advanced artificial judicial intelligence." *American Philosophical Quarterly* 60 (2): 187–203. doi.org/10.5406/21521123.60.2.06