Why does the machine punish?

The effects of the use of machine learning in criminal sentencing on the application of the theories of punishment

Student: Matthijs de Lange
Delft University of Technology
Supervisor: Evgeni Aizenberg PhD.
Responsible professor: Prof. dr. ir. Inald Lagendijk

Abstract

In recent years, the practice of risk-assessment has started to utilize machine learning to take in larger data sets and improve prediction accuracy. Simultaneously, it has expanded into the sentencing stage of the criminal justice system. This paper analyzes the effect of these developments on the consideration of the theories of punishment during sentencing. After first going over key characteristics of both the theories of punishment and the practice of risk-assessment, it presents a series of connections between aspects of machine learning enabled risk-assessment and each of the theories of punishment. It then provides developers of such systems with both general advice and advice aimed specifically at the highlighted effects, rooted in various design methodologies such as Systemic Design and Value-Sensitive Design, in order to better account for these effects. Future research can expand upon this paper both in depth and in scope: in depth by carrying out the empirical research needed to verify and improve upon the ideas in this paper; in scope by extending this research to other ethical debates surrounding machine learning in criminal justice.

Keywords

machine learning, criminal justice, risk-assessment, theories of punishment, sentencing

Introduction

Why do we punish? What gives us the right to punish those that have committed a crime? Is it simply because they deserve it, or does it serve some greater good to do so? In case of the former, what do they deserve and why? In case of the latter, what greater good does it serve and how?

Questions such as these have been the subject of debate among philosophers and political scientists since time immemorial. The answers to these questions have coalesced into various schools of thought, generally referred to as the *theories of punishment*; lawyers, lawmakers, and judges face these in a very real way during sentencing considerations.

With the introduction of machine learning (henceforth, ML) into the realm of criminal justice, these theories have now become relevant to computer scientists as well. In recent years, predictive algorithms that had thus far been used to inform decisions to release defendants on bail before trial, or to release them on parole or probation afterwards have gradually moved into the sentencing stage (Barry-Jester et al, 2015; Christin et al., 2015). There, they inform judges of the defendant's chances of recidivism, which in turn informs the nature and severity of their sentence (Barry-Jester et al, 2015)[1]. Alongside this development in their application, these algorithms have become increasingly technically advanced, more often than not through augmentation by ML (Kehl & Kessler, 2017).

This newfound use of ML in the sentencing stage of the criminal justice system thus calls for computer scientists to actively consider the theories of punishment, as the decisions made in the design and development of such ML systems now affect how the questions they address are answered in practice. The various theories differ fundamentally in which factors to consider when determining someone's punishment, so the nature and design of ML systems could lead to significant bias for one theory over another (Eaglin, 2017; Donohue, 2019; De Diego Carreras, 2020).

[1] Although only cited with one source here, this fact can be found in the majority of the literature on risk-assessment cited in this paper, citing all would result in an undesirably long citation.

It has long been recognized that the technology we use has an impact on the values we hold and express, and that it can nudge us one way or another (Winner, 1980; Friedman et al., 2008). Any tool inevitably facilitates certain uses and subsequently promotes certain norms and values, while being unsuited for and inhibiting others (Friedman et al., 2008). In response to this, various value-conscious design methodologies have been developed over the years such as Value-Sensitive Design (Friedman, 1997; Friedman and Hendry, 2019; Friedman et al., 2002), Values in Design (Flanagan et al., 2008; Nissenbaum, 2005) and Participatory Design (Schuler and Namioka, 1993). A common denominator among these methodologies is the early fostering of a thorough understanding of the social context that one is designing for, and the normative consequences that one's design might have. In some cases this might lead to the conclusion that the technology under development is ultimately not suited for the problem or context. Selbst et al. (2019) call special attention to this situation, as they note that often developers fall into the "solutionism trap" where they take for granted that the technology is in fact the right solution.

In this paper I will build upon the existing discourse to provide an in-depth research analyzing the relations between the design and development choices made by computer scientists in developing ML systems for use in sentencing, and the various theories of punishment. Subsequently, I will discuss ways forward, drawing on ideas from various of the aforementioned design methodologies to inform developers of ways to account for the various implications of their design presented in this paper. On various occasions, I will harken back to the solutionism trap, and recommend development be stopped.

In the first section of this paper I will go over the various theories of punishment. First it will consider all fundamental theories separately in order to convey the main ideas. Then it will discuss the idea of pluralistic theories and the theory of limiting retributivism in particular, as it is especially relevant to this paper.

In the second section I will discuss the practice of risk-assessment, which is ML's primary application in criminal justice. It will go over the history of this practice by means of four generations it is often divided into, describing its key aspects in the process.

In the third section I build upon the background theory presented in the first two sections, presenting a series of relations between the use of risk-assessment and ML, and the theories of punishment.

In the fourth section I then discuss the implications of these relations on the design process. Following the highly conceptual contents of the first three sections, this section will first propose the use of Value-Sensitive Design as a framework for the development of ML-based risk-assessment tools. Then it will identify the relevant stakeholders and briefly introduce how to engage said stakeholders in value-centered ways. After this, it will propose various additional ways to incorporate and build upon the knowledge presented in the first three sections in the next stages of the design and development process, utilizing ideas and methods from the aforementioned design methodologies.

Finally, I will summarize, discuss some noteworthy topics omitted from this paper due to its scope and that require more research, and conclude.

Method[2]

This research is a qualitative literature study. The literature was gathered primarily through Google Scholar, with occasional use of normal Google searches and Microsoft Academic. Sources were gathered either by keyword search or by the snowball method.

The theories of punishment

The discourse around the grand question of "why do we punish?" is old, and the body of literature surrounding it immense. As such, the full extent of this discourse is far beyond the scope of this research. That being said, a solid understanding of the most prominent theories is essential for

[2] See the "responsible research" section at the end of this paper for an ethical reflection upon my methods and research.

understanding the ideas presented in this paper. In this section I will therefore attempt to summarize as concisely as possible the relevant parts of the discourse.

Before doing so, it is important to note that theories of such complex nature should not be seen as fully distinct from one another. Gray areas and ambiguity are inevitable and so, neat packaging will only lead to a loss of nuance and arbitrary demarcation (Wood, 2010a). However, at their core, each theory does represent a fundamentally different idea, which is what I will present in the first part of this section. After having done that, I will address a popular hybrid (or pluralistic) theory, which tries to make various theories work together.

The fundamental theories

At the top-most level, the theories of punishment can be divided into two categories: consequentialist and non-consequentialist (Frase, 2005; Wood, 2010a; Wood, 2010b). Consequentialist theories justify punishment by virtue of what it achieves, generally some benefit to society as a whole; the most prominent example being the reduction of future crime (Frase, 2005; Wood, 2010a). Non-consequentialist theories justify punishment by virtue of the principles that it upholds; it is simply the right thing to do (Wood, 2010b). They argue that punishment is not a means to some greater end, but rather something we do in service of principles such as justice and fairness (Frase, 2005; Wood, 2010b). We will first look at consequentialist theories, followed by non-consequentialist ones, and lastly a theory that can be put in either category.

Consequentialist theory

As said, consequentialist theories consider only the consequences of punishment in order to justify it. The most prominent of these theories is utilitarianism, which narrows the broad idea of "consequences" down to those things that bring about the maximum amount of happiness or the greatest benefit to society (Wood, 2010a). In practice, this usually means a reduction or prevention of future crime (Frase, 2005). There are five generally accepted ways[3] to achieve this feat: *rehabilitation, incapacitation, specific deterrence, general deterrence,* and *denunciation* (Frase, 2005). I omit the lattermost, denunciation, in this section to discuss it under *communicative theory*, as it can also be considered as non-consequentialist.

Rehabilitation, incapacitation, and specific deterrence share the common feature that they are all aimed at reducing the individual's chance of reoffending (Frase, 2005). *Rehabilitation* assumes that there are identifiable, causal factors for one's criminal actions and consequently seeks to mitigate these through some form of treatment (Frase, 2005; McNeill, 2014). The nature of these causal factors can vary wildly, from the offender's socio-economic circumstances or moral character to their mental well-being or habits of substance abuse (McNeill, 2014). *Incapacitation* attempts to remove the option of re-offending from the offender's set of possible actions (Hoskins, 2019). The most commonly known and used way of doing this is by incarceration, physically removing the individual from the rest of society (Hoskins, 2019). *Specific deterrence* tries to reduce the probability of an individual reoffending by the threat of being punished in the same way that they were before, or worse (Stafford & Warr, 1993; Frase, 2005). *General deterrence* differs from the previous three in who it aims to affect. Although closely related to specific deterrence, general deterrence hopes to discourage all other members of society from committing crime through the threat of receiving the same punishment as the offender (Stafford & Warr, 1993).

Non-Consequentialist theory

Although not the only one, non-consequentialism's most prominent representative is retributivism, which sees the punishment of those who have done wrong as an intrinsically good thing (Alschuler, 2003) regardless of what it achieves (Wood, 2010b). To determine the severity of punishment, retributivism states that it should be proportional to the blame that belies the offender (Hart, 1968/2008), where blame is constructed from two factors: the harm caused by the offender's crime and the degree to which the offender is culpable for that harm (Frase, 2005). The

[3] To clear some possible confusion: I speak of "consequentialist theories" in plural because these "ways" are generally seen as theories of their own.

degree to which an offender is culpable depends on an array of factors such as mental state, intent, or situational factors that might have coerced the offender into committing the crime (Frase, 2005).

Communicative theory

The last commonly recognized, fundamental theory of punishment is the communicative theory, also referred to as expressive theory or denunciation. Its core idea is that punishment serves to express or communicate condemnation of the offender's actions (Wringe, 2017). This theory can be considered both consequentialist (Frase, 2005, Wood, 2010b) and non-consequentialist (Wood, 2010b).

From a consequentialist's perspective denunciation serves to reduce or prevent future crime by reinforcing the norms and values underlying the law (Wood, 2010a). Similar to the two types of deterrence, this can be done either through communicating to the offender themselves that what they did is wrong, or by expressing this to society at large (Wringe, 2017).

The non-consequentialist view of denunciation holds that it is an end in and of itself: "*Punishment is justified insofar as it represents, symbolizes, or gives dramatic statement to the community's deepest moral beliefs, and voices the society's abhorrence, resentment, and disgust at the crime*" (Wood, 2010b, p. 4).

Pluralistic theory

These fundamental theories are neither mutually exclusive nor independent of one another; they can be used alongside each other, reinforce each other, or interact in other ways. Various scholars have gone a step further and argue that we not only *can* use them in conjunction, but that we *should*. Some do so on theoretical or philosophical grounds, stating that no single value or idea is enough to justify punishment, and that the various values that together would be enough cannot be reduced to a singular idea (Ten & Ashby, 1987; Lacey, 1994). Others base their argument on more practical considerations, reasoning that in real world situations there are always multiple interests and ideas at play and so we should have a theory of punishment that encompasses more than a single idea (Frase, 2005). Such theories are called hybrid or pluralistic theories, and the one most prominent in practice and relevant to this paper is called "limiting retributivism" (Morris, 1974).

Limiting retributivism gives retributivism and utilitarianism two distinct roles in determining the sentence for a particular crime. The retributivist idea of what the offender deserves is used to set upper and lower bounds to the sentence they can be given, and utilitarian considerations can subsequently determine where the sentence should fall within that range (Frase, 2005). Two things make the theory of limiting retributivism of particular interest to this paper: first, it is widely adopted in practice and thereby warrants consideration. Second, it places retributivist considerations in the hands of lawmakers, outside the courtroom, disconnected from any individual case. As I further illustrate later in the paper, this allows for much better synergy between ML and retributivism. Now we move to discussing the role that ML fulfills within the criminal justice system.

The application of machine learning in criminal justice: risk-assessment

The use of ML in the criminal justice system has not yet brought us robot judges, but its integration is rather far-reaching nonetheless. Encompassing nearly all stages of the system, ML is used in policing (Završnik, 2019; Završnik, 2020), pre- and post-trial situations such as bail-setting and probation decisions (Christin et al., 2015; Harcourt, 2015; Monahan & Skeem, 2016; Kehl & Kessler, 2017), and most recently, sentencing (Berk & Hyatt, 2015; Kehl & Kessler, 2017; De Diego Carreras, 2020)[4].

[4] For a more elaborate set of sources on the spread of predictive tools in criminal justice (although not specifically based on ML) see footnote 1 in *Constructing Recidivism* (Eaglin, 2017).

In this paper I focus on the latter stage of sentencing, as it is here that the theories of punishment are applied. Here, the main purpose of ML algorithms is to predict the danger of the offender in order to inform the judge's decisions on the nature and severity of their sentence. This practice is generally referred to as "risk-assessment" (in this paper referred to as risk-assessment). Other applications of ML in the sentencing stage have been theorized, such as using ML to interpret the law (Davis, 2018), but its application to risk-assessment is by far the most prominent one, and the only one that has been widely adopted in practice.

It is important to understand however, that risk-assessment has been around far longer than ML. In the US, its use can be traced back as far as the 1920s (see Burgess, 1928) and it has played a prominent role in the justice system ever since (see Harcourt, 2005; Harcourt 2015). Throughout those years the practice of risk-assessment has seen several identifiable generations (Bonta & Andrews, 2007; Taxman et al., 2014; Kehl & Kessler, 2017), meaning that the augmentation of the procedure by ML can be better understood as the next evolution of risk-assessment than as a fundamentally new thing. In the rest of this section I will therefore elaborate on the practice of risk-assessment. I will do so by going over the aforementioned generations and the key aspects that characterize each one[5].

The first generation

The first generation of risk-assessment did not yet have much to do with the statistical approach of today. For several decades evaluating the danger of an offender was left to correctional staff such as probation officers and prison staff, or clinical professionals such as psychiatrists and social workers, making the assessment of risk a matter of professional judgement (Bonta & Andrews, 2007). This generation of risk-assessment has been widely criticized (Meehl, 1954; Grove & Lloyd, 2006), particularly for the unchecked influence of human bias (Taxman et al., 2014) and has been almost, if not fully replaced by the later generations of risk-assessment (Taxman et al., 2014).

The second generation

In the early 1970s there was a growing consensus that risk-assessment needed to shift from professional judgement to more scientific, objective measures[6] (Bonta & Andrews, 2007). The following decade consequently saw the introduction of the evidence-based or actuarial approach to risk-assessment (Bonta & Andrews, 2007) that risk-assessment is known for today. Instead of the professional judgement of the first generation, actuarial risk-assessment relies on a set of offender characteristics that have been found to indicate recidivism, which are then quantified in order to produce a risk score (Bonta & Andrews, 2007). This new approach quickly gained in popularity and became more and more widely adapted, as research continued to show its superiority over clinical or professional judgement (Bonta & Andrews, 2007; Meehl, 1954; Grove & Lloyd, 2006).

One aspect of second-generation risk-assessment tools that is important to note is that the characteristics were not chosen based on any theory or known causal relation (Bonta & Andrews, 2007; Barabas et al., 2018). Instead, correlation and availability dictated what factors were included (Bonta & Andrews, 2007). Consequentially, most of the factors were of (criminal) historical nature, as this was most plentifully available to judicial institutions (Bonta & Andrews, 2007; Barabas et al., 2018). This meant that almost all factors were of a static, immutable nature and could therefore not account for any change in an offender's behavior or nature (Bonta & Andrews, 2007). This made the second-generation risk-assessment tools unfit for informing treatment decisions (Gendrau et al., 1996; Kehl & Kessler, 2017).

The third generation

To remedy the shortcomings of the second generation, third-generation risk-assessment tools started to incorporate more dynamic factors (Bonta & Andrews, 2007; Taxman et al., 2014;

[5] A very similar section can be found in *the moral (un) intelligence problem of artificial intelligence in criminal justice: a comparative analysis under different theories of punishment* (De Diego Carreras, 2020). The differences between our papers do not warrant a difference in these sections, making the similarity unfortunate but unavoidable.

[6] Although it came to fruition in the 1970s, the idea that there could be measurable predictors for recidivism had been around much longer (Burgess, 1928; Taxman et al. 2014).

Barabas et al., 2018). Dynamic factors are characteristics of an offender that are not anchored to their past but rather those that describe their current situation and can be changed or treated (Gendrau et al., 1996). Examples include mental illness, criminal peers, or substance abuse (Gendrau et al. 1996). Due to the presence of these amendable factors in the assessment, which are often referred to as "criminogenic needs" (Taxman et al., 2014; Kehl & Kessler, 2017; Barabas et al., 2018), third-generation risk-assessment tools are generally referred to as "risk-need" tools (Bonta & Andrews, 2007). These risk-need tools are far better suited to rehabilitation efforts, given the ability to evaluate the change in risk score based on a change in the dynamic factors (Bonta & Andrews, 2007). Additionally, the predictive power of dynamic factors has been shown to be at least as good as that of static factors (Gendrau et al., 1996).

The fourth generation

The fourth and most recent generation of risk-assessment is often seen as an extension of the third (Taxman et al., 2014), as there are no major changes in *how* fourth-generation tools reach their conclusions. Rather, fourth generation tools differ in their focus on matching offenders to an appropriate treatment instead of simply outputting a risk score (Taxman et al., 2014). That being said, in order to achieve these improved results effectively and due to these tools naturally improving over time, the set of factors that is being considered has significantly expanded as well (Bonta & Andrews, 2007). It is also in this latest generation of risk-assessment that we see more technically advanced tools and specifically, the use of ML to create tools that can take in ever increasing amounts of data and adapt their predictive models over time (Kehl & Kessler, 2017; Donohue, 2019).

In addition to this generational overview, it is important to appreciate that the use of risk-assessment in sentencing is a relatively recent development (Starr, 2014; Christin et al., 2015). Its original and primary use throughout most of its century-long lifetime has been in parole and probation decisions (Harcourt, 2005; Casey et al., 2011, Christin et al., 2015), and although more recent than parole and probation, risk-assessment has also already been widely adopted into the pretrial stage (Christin et al., 2015; Dalakian, 2018). The scope of this paper precludes me from considering all questions one could raise about this development. Instead, in the next section I will outline how the various theories of punishment will be affected by this newfound use.

The effects of machine learning and risk-assessment on the theories of punishment

Sentencing is a wildly complex endeavor, perhaps the most complicated task a judge faces in their professional capacity (Donohue, 2019). It warrants consideration of a myriad of factors, not the least of which is the ultimate purpose of the sentence. The rising use of risk-assessment, and with it ML, in this process is all but guaranteed to impact these considerations. In this section I will present a series of effects that the use of ML-enabled risk-assessment in sentencing could have on the consideration of the theories of punishment during the sentencing process.

First, I present two theory-independent effects which do not affect a particular theory, but work on a more meta-level, affecting the role of the theories of punishment in a more general way. Then I present a series of theory-specific effects, each going over how the use of ML-enabled risk-assessment affect a particular theory or set of theories. These will be presented loosely in order of least compatible to most compatible.

Codifying implicit considerations

As I highlighted in the first section of this paper, there has long been much discussion on the theories of punishment. It is undeniably a topic of great philosophical interest. In practice, however, its importance might not be so apparent. This is because in day-to-day proceedings, such considerations are generally left implicit. Judges need not, and often do not, fully or exactly disclose the factors they considered in order to reach their decision (Donohue, 2019). With the rise of ML however, these considerations must be written into code, something which unavoidably requires them to be made explicit (Coyle & Weller, 2020). Moreover, following the controversies around

ML systems in criminal justice and the (racial) biases they propagate (see e.g. Angwin et al., 2016), many have called for such systems to be made more transparent. This means that not only do these considerations have to be made explicit at the creation of such a system, they will most likely remain visible as such ever after (Coyle & Weller, 2020).

"The machine as manipulator"[7]

Although speculated about (see De Diego Carreras, 2020), ML-based risk-assessment tools will most likely not be making any autonomous sentencing decisions in the near future. However, although not as dramatic, these tools still propagate the biases they have regarding the theories of punishment through the people that use them, this being the sentencing judges. One way in which this can happen is a process called "anchoring", where someone bases their decision mainly on the first piece of evidence that is presented to them (Starr, 2014; Christin et al., 2015; Donohue, 2019). Thus, when a judge is consistently presented with a risk score at the start of a sentencing hearing, they might unknowingly adjust their decision to match said score (Christin et al., 2015). Moreover, if the risk-assessment tool providing that score is biased towards for example, incapacitation, the judge might unconsciously adapt or propagate that bias (Donohue, 2019).

Another well-known way in which these tools can distort the user's thinking is through automation bias. Judges, usually not trained in computer science or statistics, are unlikely to challenge such a technically sophisticated process; the scientific, numerical outputs that ML-based risk-assessment tools produce appear more objective than one's own judgement (Christin et al., 2015). Consequently, it is rare for judges to override the tools even despite their own rich legal expertise (Christin et al., 2015).

Starr (2014) suggests that the very presence of risk-assessment tools can serve to indicate to a judge that an offender's risk should be a significant factor in their sentencing considerations, an idea supported by the fact that in at least 4 US states the use of risk-assessment during sentencing has been made a requirement or is strongly recommended (Monahan & Skeem, 2016; Eaglin, 2017).

Non-consequentialism: no use for prediction

Non-consequentialism has no need for any sort of predictive capacity given that non-consequentialism, by definition, is not concerned with the future when motivating an offender's punishment (Starr, 2014; De Diego Carreras, 2020). Both retributivism and non-consequentialist expressivism base their motivation of punishment exclusively on the offender's past and both draw their conclusions based predominantly on moral considerations (Frase, 2005; Wood, 2010b; De Diego Carreras, 2020). This backward-looking perspective and reliance on morality make risk-assessment and/or ML essentially unfit to serve non-consequentialist goals, as they are respectively forward-looking by definition and incapable of moral thoughts as of yet (Davis, 2018; De Diego Carreras, 2020). Conversely, this means that the increasing prevalence of such predictive tools leaves increasingly less room for non-consequentialist considerations.

Notwithstanding the above argued effect, retributivism and ML-based risk-assessment might still find themselves used alongside one another. The theory of limiting retributivism gives retribution and risk-assessment separate roles, allowing room for both in the process. Under limiting retributivism retributivist considerations are captured in (case-independent) guidelines for upper and lower sentencing limits, ensuring no unjust punishment; risk-assessment can subsequently influence where within that range the actual punishment falls (Slobogin, 2019). This scenario thus tempers the antithesis between ML-based risk-assessment and retributivism somewhat, but this solution is still incompatible with pure retributivism, removes retributivism from case-by-case considerations, and (depending on how loose or tight the limits are set) could still lead towards an overall shift away from retributivism.

Deterrence and expressivism: no use for recidivism

Although the use of ML-based risk-assessment is unfit for non-consequentialist goals, this does not mean it serves all consequentialist goals well. Both specific and general deterrence, as well as

[7] This heading is a quote from (Donohue, 2019)

consequentialist expressivism have no need for the use of ML-based risk-assessment tools, as none of these goals benefit from knowledge regarding an individual offender's risk of future crime (Monahan & Skeem, 2016). Although they do base their sentencing rationales on the future effects of punishment, the issue lies in the intended effect and for the latter two, the target as well. Basing an offender's punishment on considerations of instilling fear or moral values, means that an offender's chances of recidivism play little to no role (Monahan & Skeem, 2016; Kehl & Kessler, 2017). Additionally, both general deterrence and some forms of consequentialist expressivism try to affect society at large, rather than the individual offender currently being sentenced (Frase, 2005), which renders any considerations of the individual offender's future irrelevant. Moreover, the population-wide effects that these two consequentialist theories hope to achieve are of such complex, long-term and indirect nature that it is unlikely any computer system can offer useful predictions for achieving them, even if it was not in the context of risk-assessment.

What is important to realize however, is that these theories can be considered in tandem with others. A sentence that incapacitates an individual can also deter the general public; a sentence that rehabilitates can still express that the offender committed a morally wrong act. However, an appropriate theoretical framework is needed to enable proper coexistence. Such a framework would guide decision making, securing a role for deterrence and/or expressivism, and allowing intentional consideration of these theories and preventing them from being pushed to the side.

Rehabilitation: dynamic vs. static data

The phrase "garbage in, garbage out" rings familiar to many computer scientists and effectively captures the idea that within the realm of ML the input data is possibly the most important factor in determining an algorithm's behavior. This is an important consideration for all developers of ML systems, and one most are well aware of when it comes to ensuring the technical functionality of their system. Yet, as Eaglin (2017) points out, the normative considerations implicated in selecting and processing data are often neglected.

Decisions regarding the data selected by the developers of ML-enabled risk-assessment tools such as COMPAS (see e.g. Angwin et al., 2016)[8] are currently driven by mathematical accuracy and availability (Eaglin, 2017). Although the need for accuracy does not necessarily force the use of static data (Gendrau et al., 1996), the focus on availability certainly does, as became clear in the second generation of risk-assessment tools (Bonta & Andrews, 2006; Eaglin, 2017). This has led many risk-assessment tools currently in use to lean heavily on criminal history, rendering them unfit for rehabilitative purpose once again, like the second-generation tools before them (Bonta & Andrews, 2006). Thus, even if rehabilitative ideals are not fundamentally unsuited for consideration by ML-based risk-assessment tools, their cooperation requires the normative aspect of data selection to be considered at the development of these tools, something that might go against the economic incentives of the developers.

Incapacitation: the perfect fit

So far, we have seen tensions between almost all theories of punishment and the use of ML-based risk-assessment and although none of them appear insurmountable as of yet, it will require conscious, concentrated effort to do so. The one exception to this is the theory of incapacitation. The rationale of incapacitation is the one theory of punishment that is most fully in line with the predictive sentencing promoted by the use of ML-based risk-assessment tools (Harcourt, 2015; Hoskins, 2019; Eaglin, 2017). If the sole purpose of punishing someone is to keep them from committing another crime, the risk of them doing so is naturally the most important factor to consider in determining their sentence. Many of the issues that arise in relation to the other theories also quickly wither away; incapacitation is neither backward looking nor morally motivated like retributivism, its goal has a logical connection with recidivism risk unlike deterrence and denunciation, and accounting for change in an offender is of far lesser importance to incapacitative theory than it is to rehabilitative theory.

[8] For a more general overview of this particular software see (COMPAS (software) - Wikipedia, n.d.)

In summary, the only theory of punishment fundamentally in-line with ML-based risk-assessment is incapacitation. For retributivism, prediction is of no use in general; for expressivism and deterrence, predictions could be useful, yet recidivism is not relevant; and for rehabilitation the type of data used determines compatibility, as assessment needs to be based on dynamic data in order to allow for rehabilitative considerations.

The use of ML-based risk-assessment bring about two more, theory-independent effects. Namely, that (1) the choice for a certain theory of punishment needs to now be made explicit in order to be incorporated into a tool or put into code and (2) that tools could strongly steer judges towards whatever theory that tool might be biased towards, through "anchoring" and automation bias.

Implications for socio-technical design

The effects and implications discussed in the previous section have not presented as, nor should they be seen as things to be avoided per se. Preferences for one theory or another differ, understandably, from culture to culture and from person to person. That being said, even if one agrees with the bias towards incapacitation that the use of ML-based risk-assessment tools imposes on the system, it is important to be aware and in control of them. However, most design decisions for these tools are currently being made without any awareness of their normative implications (Eaglin, 2017).

In this section I will build upon the conceptual work of the previous one, proposing steps that can enable developers to account for the presented effects. These steps include the use of certain design methodologies, specific ways to interact with or questions to ask the various stakeholders, and it will detail who those stakeholders are.

First, I will discuss the use of Value-Sensitive Design, as it provides a framework for the more specific measures I propose. Then, the relevant stakeholders will be put forth and lastly, I will address the effects from the previous section, for each proposing how to account for them, grouping effects together if they benefit from similar measures.

Value-Sensitive Design

To accommodate all the measures presented in the rest of this section, I first propose that development of risk-assessment tools utilize an overarching methodology like Value-Sensitive Design, which I will elaborate on here, or something similar such as Values in Design (Flanagan et al., 2008).

The term Value-Sensitive Design was first coined in 1996 by Batya Friedman (Friedman, 1996) and has since evolved into a well-established design methodology aimed at incorporating human values into the design process (Friedman et al., 2002; Cummings, 2006; Friedman et al., 2008). Its core idea exists of a tripartite structure through which to organize the design process: conceptual investigations, empirical investigations, and technical investigations (Cummings, 2006, Friedman et al., 2008). Although the order of these is logical, it is equally important to recognize that it is not absolute. The three types of investigations represent an iterative and integrative process, where each phase informs the others, and each phase is visited repeatedly throughout the development cycle (Friedman et al., 2013).

Conceptual investigations are meant to explore what values and stakeholders are implicated by the technology under development, and to define the values in question (Friedman et al., 2008). The first three sections of this paper, along with the "Stakeholders" section hereafter fall squarely under conceptual investigations. *Empirical investigations* aim to validate, update, and deepen the findings from the conceptual investigations (Friedman et al., 2008). Through quantitative and qualitative methods from the social sciences, empirical investigations should try to find how the values and concepts are applied in practice (Friedman et al., 2008). This can include steps such as observation of stakeholders, asking them about past (value-laden) experiences and user-centered prototyping. *Technical investigations* can seem much like the empirical ones but focus specifically on how the values translate into the technology and how the limitations of the technology translate into value limitations (Friedman et al., 2008).

Due to its iterative nature, Value-Sensitive Design might work well in conjunction with something like Agile software development (Beck et al., 2001), augmenting it to accommodate consideration of normative and ethical implications (Aizenberg & van den Hoven, 2020). The viability of cooperation between these two methodologies should be an interesting topic for future research.

Stakeholders

Engaging properly with the relevant stakeholders is paramount to any value-conscious design methodology. Value-Sensitive Design's *empirical investigations*, Values in Design's equivalent *empirical mode* and the entirety of Participatory Design revolve around this idea. (Friedman et al., 2008; Flanagan et al., 2008; Schuler and Namioka, 1993). Although stakeholder engagement is important in non-value-conscious design methodologies too, the terms "properly" and "relevant" take on new meaning when used in a value-oriented context.

Properly engaging with stakeholders means not only discussing the classic set of functional and non-functional requirements of a project, but also engaging in conversation with stakeholders on the norms and values they hold and how these should be incorporated in, and how they are impacted by the technology under development (Friedman et al., 2008). *Relevant* stakeholders now include both direct and indirect stakeholders. Direct stakeholders include those people directly involved with the system, generally clients and users (Friedman et al., 2008). Indirect stakeholders are those people that do not directly interact with the system, but are affected by it (Cummings, 2006; Friedman et al., 2008)

Given these definitions, there are three relevant stakeholders that should be considered in value-conscious development of a ML-based risk-assessment system: lawmakers and judges as direct stakeholders, and people of the society or community that the system will be deployed in as indirect stakeholders.

Lawmakers

Lawmakers can be seen as the clients, as they represent the institutions that risk-assessment tools are generally commissioned for, such as national governments, provincial or state governments, or even more local jurisdictions. Moreover, they control the legal system in which the risk-assessment tools are being implemented. They, for example, are the ones who can change sentencing guidelines if the implementation of a risk-assessment tool requires it.

Judges

Judges represent the users in our context, they are ultimately the ones using the risk scores presented by the risk-assessment tools in their decision-making.

The people

The group of stakeholders that we only find when considering a value-conscious development is the people of the community or society to which the risk-assessment tool will be introduced. It is the people that will be assessed, and moreover, the people who decide the norms and values of their own society, which a risk-assessment tool should adhere to.

Two particular subgroups should be considered in particular: Defendants and public prosecutors. Defendants make up the subset of the people that has been or is most directly subjected to risk-assessment and consequent punishment. They can thus bring first-hand experience to the table on the practical realization of the different theories of punishment. Public prosecutors warrant consideration as they represent the people within the legal system, while also possessing a great deal of legal expertise. They could thus prove instrumental in translating community values into legal terms.

Implicit to explicit: stakeholder engagement

As I described in the section "Codifying implicit considerations", the development of risk-assessment tools will inevitably make explicit values previously held implicitly. Besides the implications mentioned in that section, it also means that during development, stakeholders will have to go through the same transition. Thus, it will take a conscious effort to elicit usable knowledge on stakeholders' positions on the theories of punishment as they might not be fully aware of it themselves. Lawmakers and judges will most likely be familiar with the topic, making things somewhat easier, but especially the general public could have issues properly articulating their opinions. I propose two ways to help ensure constructive communication of values between the stakeholders and developers.

First, interviews with stakeholders should contain plenty of indirect and in-depth lines of questioning (Friedman et al., 2008). Direct questions such as "what theory of punishment do you prefer" can seem efficient and to the point, yet often do not yield the best results. Within the context of the implicit-explicit transition this is even more so the case. Stakeholders might not be familiar with the terms, and even if they are, they might not have considered this question in an explicit manner before. Indirect questions such as "Do you think people can change for the better" mitigate these issues as such questions are generally more familiar or intuitive. The developers can subsequently link the questions and answers to a theory of punishment. Additionally, it is important to make sure to go in-depth, asking "why" whenever possible or relevant. This not only deepens the developer's understanding of a stakeholder's motivation (Friedman et al., 2008), it might also help the stakeholder to articulate their position, something they might especially struggle with if they had not explicitly done so before.

The second method I propose is an adaptation of value scenarios (Nathan at al., 2007)[9]. Namely, in order to get a solid grasp on a stakeholder's preferred theory of punishment I propose to present them with concrete offender cases. By presenting the stakeholders with anonymized or entirely fictional cases and asking them to choose a certain sentence for the imagined offender, they can choose to do so fully based on intuition, leaving the transition to an explicit theory of punishment in the hands of the developers. If more information is needed or wanted, stakeholders can always be asked to explain their reasoning afterwards, which is also made easier by the fact that they can now base their explanation on a concrete example.

Systemic Design

Systemic design incorporates the ideas of systemic thinking into a design methodology. Systemic thinking is based around the idea of *synthesis,* where properties of a whole (a system) only emerge when its parts come together (which is directly opposed to *analysis*, which hopes to find the properties of the whole by studying the parts) (Van der Bijl & Malcolm, 2020). Consequently, systemic design holds that designers and developers should look beyond only their own product and account for the interconnectedness of the (social) system they are bringing it into (Van der Bijl & Malcolm, 2020). Given the nature of the following set of issues, I propose the use of Systemic Design in the development of risk-assessment tools, or at least highly recommend the consideration of systemic thinking.

Namely, each of the following issues arises not from any particular aspect of the risk-assessment tool itself, but rather from its relation to the judge or the time and place of use within the sentencing process, in short, the way it is integrated into the criminal justice system. The measures to mitigate these issues are consequently aimed at the system around the tool rather than at the tool itself. During development lawmakers and judges should be engaged in order to ensure that the finished risk-assessment tool is integrated into a system that is suited for it and that it supports the value requirements of all stakeholders.

Accounting for human-computer interaction

In the "machine as manipulator" section I described how judges can be excessively influenced by the presence and use of risk-assessment tools during sentencing. This had three distinct causes: first, the psychological process of "anchoring", where the first piece of evidence "anchors" a judge to a certain decision. Second, automation bias, where the machine is trusted over one's own judgement due to the fact that it is (perceived as) a sophisticated, objective machine. Third, the

[9] I based this idea off value scenarios, but I am aware that it bears similarity to various other methods.

signaling effect that the presence of a risk-assessment tool has, reminding judges that recidivism risk should be taken into account.

To combat the anchoring issue developers should engage in targeted user testing with various judges to assess how the effect of a risk-assessment tool on a judge changes in relation to the tool's place within the process. If the risk score is presented at different times during a judge's decision-making process, anchoring might occur to a lesser degree. Additionally, training judges against anchoring might reduce the effect as well (Christin et al., 2015).

Similarly, automation bias could also be (partly) mitigated through thorough training of judges in the use of such tools. If judges understand better how these tools come to their conclusions and how they are supposed to interpret the produced scores, they might be more comfortable going against it or questioning it.

Lastly, the signaling effect's strength is determined primarily, if not exclusively, by the prominence that the jurisdiction gives to the risk-assessment tool. If, like in the aforementioned states, its use is made mandatory, it stands to reason that the signaling effect will be greater. This can and should be further explored through user testing with judges.

Although the implementation of these measures will require systemic design, as mentioned, the nature of the measures revolves around human-computer interaction. I therefore recommend involving an human-computer interaction expert in the process.

Accounting for the legal system

As I showed in the sections "no use for prediction" and "no use for recidivism" risk-assessment tools are fundamentally incompatible with retribution, deterrence and communication, consequentialist or not. For these theories to coexist with risk-assessment, changes to the risk-assessment tool itself will not help, instead it is the legal system this tool is implemented into that should be adjusted to provide space for these theories to be considered. Already mentioned was the idea of using the pluralistic "limiting retributivism" theory to create some degree of coexistence for retributivism.

This creates two relevant scenarios[10]: The first is that stakeholders indicate that retributivism, deterrence, or communication should remain an important factor in sentencing decisions, but do not hold it as their primary or sole objective. The second scenario is that stakeholders indicate that they hold these theories as their primary or sole objective, or believe they should be applied on a case-by-case basis.

In the first scenario, in the case that the particular theory is retributivism, limiting retributivism can provide a way forward for the project. Developers should engage with lawmakers in order to assess whether limiting retributivism can be adopted in their jurisdiction. If it is already in place, they should discuss if and how it should be modified to suit the introduction of the risk-assessment tool. In case it is either of the other two theories, the way forward is similar, although possibly more laborious. These theories do not yet have such well-established theories, and so these will have to be developed as part of the development process.

In case that any of these theories is the stakeholders' primary or sole motivation for punishment, the solution is straightforward. This means a ML-based risk-assessment tool is simply unfit for the stakeholders' requirements. Avoiding the solutionism trap (Selbst et al., 2019), development should be terminated.

Data-conscious design

The most technically concrete relation I have presented in this paper is that between the type of data used – dynamic or static – and the suitability of the resulting system for rehabilitative ideals. What is left for developers to do is communicate this relation to the stakeholders and subsequently engage in a constructive discussion on how to deal with this relation. Most likely, the discussion will revolve around the conflict between the possible need, if stakeholders have indicated so, for the incorporation of rehabilitative ideals and the increased cost that the use of dynamic data brings with it. Developers should avoid framing this as a binary choice though, as a tool can

[10] The third natural scenario, where these theories play no role is not relevant for this section, as it considers how to incorporate them if needed.

adopt various degrees of dynamic data, against various degrees of extra costs (Friedman et al., 2008).

Discussion

In recent years, machine learning has started being used for risk-assessment and risk-assessment tools have started being used for making sentencing decisions. Consequently, an offender's risk of recidivism has gained increased prominence in sentencing, at the expense of the theories of punishment that are not concerned with this risk; retribution, deterrence and expressivism all find themselves pushed to the side by the newfound use of risk-assessment. Between the two theories that do consider recidivism risk (rehabilitation and incapacitation) rehabilitation requires the use of dynamic data, which is prone to incur extra costs. So, the use of risk-assessment tools at sentencing ultimately promotes incapacitation to the primary theory of punishment.

In order to ensure that these tools serve our norms and values instead of shaping them, I propose the use of various design methodologies to mitigate or account for these effects. Most importantly I recommend the use of Value Sensitive Design and Systemic Design to respectively incorporate value-considerations into all parts of development and to account for effects not solvable from within the risk-assessment tool alone.

The scope of this paper has kept me from discussing various other issues related to the introduction of ML and risk-assessment into the sentencing stage of the criminal justice system. Additionally, it was not within my capacity to engage in any empirical research, carrying out the steps outlined in this paper. This section will briefly go over some of the topics and steps not covered here in order to contextualize my conclusions and recommend future research.

The most prominent debate not included in this paper is that of the unfair discrimination many feel results from the use of ML to make sentencing decisions (see e.g. Angwin et al., 2016). From this, many smaller debates have sprawled, some arguing about what constitutes a "fair" system and thus when something is "unfair" discrimination (Angwin et al., 2016) others stating that such systems are fundamentally unconstitutional and that they legitimize racism, sexism, and the likes (Starr, 2014). However, scholars like Slobogin (2018) make convincing arguments that humans are not much different and, in any case, are more biased than the algorithms might be. The conclusions and recommendations in this paper should certainly be seen in context of this debate as well: the values discussed here are not the only ones to be considered in the development of a risk-assessment tool, nor are the recommendations presented here only applicable to these values. Value-Sensitive Design is a broad framework and should be applied to all relevant ethical debates surrounding the development and design process.

Additionally, Value Sensitive Design is an iterative process (Friedman et al., 2008). Not only does the conceptual research inform the empirical and technical phases that come after, the empirical and technical phases also inform and update the conceptual research. One might define certain values in a certain way during the conceptual phase but find that these are not workable definitions in practice. This means that the recommendations and implications presented in this paper might change in similar way. With this paper I hope to motivate those in a position to do so to carry out this process, updating the ideas presented in this paper accordingly.

Lastly, I have stated that specific deterrence has no use for prediction, yet as various commenters have asked there is an intuitive reasoning that it would have such a use. Namely, that if an offender is more prone to recidivism, they should be punished harsher in order to scare them off more thoroughly. Although this reasoning crossed my mind during this research and makes sense to me as well, multiple of my sources stated confidently that deterrence has no use of recidivism prediction. It was not within my capacity to build a thorough enough argument to challenge that here, but I strongly encourage future research into this topic.

Conclusions

In this paper, I have presented a series of implications that the use of ML-based risk-assessment tool in sentencing might have on the application of the theories of punishment, followed by various proposals for methods to account for these implications. It has shown that currently, the use of risk-assessment at sentencing will promote incapacitation to the primary theory of punishment. In order to account for stakeholders' positions on these matters I have proposed the use of Value-Sensitive Design and Systemic Design in order to accommodate other, more theory specific approaches. In addition to these measures I have repeatedly warned for the solutionism trap, as described by Selbst et al. (2019): if any theory besides incapacitation should be stakeholders' sole rationale for punishment no method or measure can align the use of risk-assessment with their values. Conversely the same holds: if incapacitation is a theory of punishment stakeholders are fundamentally opposed to, risk-assessment should not be the way forward.

Criminal justice is deeply rooted in society, and society's values are deeply rooted in criminal justice. With the introduction of digital technology into this domain, developers of these technologies should be aware of this. Besides the specific implications and methods presented in this paper, I hope that it has contributed to that awareness.

Responsible research

In the methodology section at the start of this paper I explained that this study relied fully on literature, and I detailed how this literature was found. This section highlights two biases in my research and acknowledges which sections of this paper are my own views, rather than direct findings.

First, during literature research, I excluded no particular authors, groups of authors or institutions. Literature was selected on the basis of whether it could provide useful information to further my research and build toward a conclusion. This was judged based on its title, abstract, and if needed, a scan of its conclusion section. Despite this, the majority of the literature ended up being from the United States, as this was more readily available. I could have tried to counter this, but I believe this would have harmed the quality of my research. Although the broader perspective could have certainly added value to the findings of my research, due to the research's time-constraints I felt I had no room to reject valuable literature nor time to spend on finding less readily available literature from other countries.

Second, I personally hold a bias towards rehabilitation. I held this opinion before starting this paper and although more nuanced now, the bias stands. In this paper, I have attempted to present all effects found in a neutral manner and address all effects equally.

Lastly, I believe it is important to acknowledge that the section "Implications for sociotechnical design" contains my own ideas on how to move forward based on the knowledge presented. This is in contrast to the previous sections, where I presented compiled knowledge found in scientific literature.

References

Aizenberg, E., & van den Hoven, J. (2020). Designing for human rights in AI. Big Data & Society, 7(2), 2053951720949566.

Alschuler, A. W. (2003). The changing purposes of criminal punishment: A retrospective on the past century and some thoughts about the next. The University of Chicago Law Review, 70(1), 1-22.

Andrews, D. A., & Bonta, J. (1995). The level of service inventory-revised. Toronto, Canada: Multi-Health Systems.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. ProPublica, May, 23(2016), 139-159.

Barabas, C., Virza, M., Dinakar, K., Ito, J., & Zittrain, J. (2018, January). Interventions over predictions: Reframing the ethical debate for actuarial risk-assessment. In Conference on Fairness, Accountability and Transparency (pp. 62-76). PMLR.

Barry-Jester, A. M., Casselman, B., Goldstein, D., Conlen, M., Fischer-Baum, R., & Rossback, A. (2015). Should prison sentences be based on crimes that haven't been committed yet. FiveThirtyEight.

Beck, K., Beedle, M., Van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., ... & Thomas, D. (2001). Manifesto for agile software development.

Berk, R., & Hyatt, J. (2015). Machine learning forecasts of risk to inform sentencing decisions. Federal Sentencing Reporter, 27(4), 222-228.

Bonta, J., & Andrews, D. A. (2007). Risk-need-responsivity model for offender assessment and rehabilitation. Rehabilitation, 6(1), 1-22.

Burgess, E. W. (1928). Factors determining success or failure on parole. The workings of the indeterminate sentence law and the parole system in Illinois, 221-234.

van der Bijl-Brouwer, M., & Malcolm, B. (2020). Systemic Design Principles in Social Innovation: A Study of Expert Practices and Design Rationales. She Ji: The Journal of Design, Economics, and Innovation, 6(3), 386-407.

Casey, P. M., Warren, R. K., & Elek, J. K. (2011). Using offender risk and needs assessment information at sentencing: Guidance for courts from a national working group. National Center for State Courts.

Christin, A., Rosenblat, A., & Boyd, D. (2015). Courts and predictive algorithms. Data & CivilRight.

Dalakian, G. J. (2018). Open the Jail Cell Doors, HAL: A Guarded Embrace of Pretrial Risk-assessment Instruments. Fordham L. Rev., 87, 325.

Davis, J. P. (2018). Law without Mind: AI, Ethics, and Jurisprudence. Cal. WL Rev., 55, 165.

De Diego Carreras, A. (2020). THE MOrisk-assessmentL (UN) INTELLIGENCE PROBLEM OF ARTIFICIAL INTELLIGENCE IN CRIMINAL JUSTICE: A COMPArisk-assessmentTIVE ANALYSIS UNDER DIFFERENT THEORIES OF PUNISHMENT. UCLA Journal of Law & Technology, 25(1).

Donohue, M. E. (2019). A Replacement for Justitia's Scales: Machine Learning's Role in Sentencing. Harv. JL & Tech., 32, 657.

Eaglin, J. M. (2017). Constructing recidivism risk. Emory LJ, 67, 59.

en.wikipedia.org. (n.d.) COMPAS (software) - Wikipedia. [online] Available at: <https://en.wikipedia.org/wiki/COMPAS_(software)> [Accessed 12 June 2021].

Flanagan, M., Howe, D. C., & Nissenbaum, H. (2008). Embodying values in technology: Theory and practice. Information technology and moral philosophy, 322.

Frase, R. S. (2005). Punishment purposes. Stan. L. Rev., 58, 67.

Friedman, B. (1996). Value-sensitive design. interactions, 3(6), 16-23.

Friedman, B. (Ed.). (1997). Human values and the design of computer technology (Vol. 72). Cambridge University Press.

Friedman, B., Kahn, P., & Borning, A. (2002). Value sensitive design: Theory and methods. University of Washington technical report, (2-12).

Friedman, B., Kahn, P. H., & Borning, A. (2008). Value sensitive design and information systems. The handbook of information and computer ethics, 69-101.

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. In Early engagement and new technologies: Opening up the laboratory (pp. 55-95). Springer, Dordrecht.

Friedman, B., & Hendry, D. G. (2019). Value sensitive design: Shaping technology with moral imagination. MIT Press.

Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works!. Criminology, 34(4), 575-608.

Grove, W. M., & Lloyd, M. (2006). Meehl's contribution to clinical versus statistical prediction. Journal of Abnormal Psychology, 115(2), 192.

Hart, H. L. A. (2008). Punishment and responsibility: Essays in the philosophy of law. Oxford University Press. (Original work published 1968)

Harcourt, B. E. (2005). Against prediction: Sentencing, policing, and punishing in an actuarial age. U of Chicago, Public Law Working Paper, (94).

Harcourt, B. E. (2015). Risk as a proxy for race: The dangers of risk-assessment. Federal Sentencing Reporter, 27(4), 237-243.

Hoskins, Z. (2019). Against incapacitative punishment. Predictive Sentencing: Normative and Empirical Perspectives, 89.

Kehl, D. L., & Kessler, S. A. (2017). Algorithms in the criminal justice system: Assessing the use of risk-assessments in sentencing.

Lacey, N. (1994). State punishment: political principles and community values. Psychology Press.

McNeill, F. (2014). Punishment as rehabilitation.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.

Monahan, J., & Skeem, J. L. (2016). Risk-assessment in criminal sentencing. Annual review of clinical psychology, 12, 489-513.

Morris, N. (1974). The future of imprisonment (p. 59). Chicago: University of Chicago Press.

Nissenbaum, H. (2005). Values in technical design. Encyclopedia of science, technology, and ethics, 66-70.

Schuler, D., & Namioka, A. (Eds.). (1993). Participatory design: Principles and practices. CRC Press.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019, January). Fairness and abstraction in sociotechnical systems. In Proceedings of the conference on fairness, accountability, and transparency (pp. 59-68).

Slobogin, C. (2019). A Defense of Modern Risk-Based Sentencing. Predictive Sentencing: Normative and Empirical Perspectives, 107.

Stafford, M. C., & Warr, M. (1993). A reconceptualization of general and specific deterrence. Journal of research in crime and delinquency, 30(2), 123-135.

Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. Stan. L. Rev., 66, 803.

Taxman F.S., Caudy M., Maass S. (2014) Actualizing Risk-Need-Responsivity. In: Bruinsma G., Weisburd D. (eds) Encyclopedia of Criminology and Criminal Justice. Springer, New York, NY.

Ten, C. L., & Ashby, C. (1987). Crime, guilt, and punishment: A philosophical introduction. Oxford: Clarendon Press.

Winner, L. (1980). Do Artifacts Have Politics?. Daedalus, 109(1), 121-136.

Wood, D. (2010a). Punishment: consequentialism. Philosophy Compass, 5(6), 455-469.

Wood, D. (2010b). Punishment: nonconsequentialism. Philosophy Compass, 5(6), 470-482.

Wringe, B. (2017). Rethinking expressive theories of punishment: why denunciation is a better bet than communication or pure expression. Philosophical Studies, 174(3), 681-708.

Završnik, A. (2019). Algorithmic justice: Algorithms and big data in criminal justice settings. European Journal of Criminology, 1477370819876762.

Završnik, A. (2020, March). Criminal justice, artificial intelligence systems, and human rights. In Erisk-assessment Forum (Vol. 20, No. 4, pp. 567-583). Springer Berlin Heidelberg.