Artificial Intelligence and Sentencing: Humans against Machines

Wingerden, S.G.C. van; Plesničar, M.M.; Ryberg, J.; Roberts, J.V.

12

Artificial Intelligence and Sentencing

Humans against Machines

Sigrid van Wingerden and Mojca M. Plesničar

12.1 Introduction

According to Chiao in his contribution to this book, the desirability of the use of AI in sentencing should be evaluated by comparing computers to the status quo ante, rather than to an unrealistic, and in any case unrealized, ideal. Although we agree that changes to the legal process such as adopting algorithmic sentencing methods can be beneficial when the change is an incremental improvement over the status quo, in order to assess whether the change is an improvement, we need to know what this "ideal" is toward which improvements are aimed. Therefore, the question whether AI is better at making sentencing decisions than human judges is approached differently in this chapter. We compare human with AI judges by evaluating the extent to which they are able to make a legitimate sentencing decision: Is legitimacy better achieved by machine than by human judges?

To answer this question, we developed a theoretical model for a legitimate sentencing regime. As explained later in the chapter, this model comprises both normative and empirical legitimacy, wherein normative legitimacy contains different levels and empirical legitimacy separate components. We use this model to compare the capabilities of AI and human judges. In reviewing human judges' performance, we draw upon research into the nature of human decision-making at sentencing. In doing so, we do not distinguish between different systems of sentencing, but rather look at universal decision-making dynamics. Regarding "AI judges," we first need to explain what we mean by the term AI judges.

Sigrid van Wingerden and Mojca M. Plesničar, *Artificial Intelligence and Sentencing* In: *Sentencing and Artificial Intelligence*. Edited by: Jesper Ryberg and Julian V. Roberts, Oxford University Press. © Oxford University Press 2022.
DOI: 10.1093/oso/9780197539538.003.0012

12.1.1. AI Judges

Defining AI is challenging. There is a myriad of definitions, and they are not static, but rather ever evolving. Generally, AI involves technology or methods that tackle problems which require intelligence to be solved (Plant 1994). Simple rule-based algorithms are usually insufficient to place in this category; the problem-solving needs to require some sort of autonomy on the side of the agent (e.g., EU-Commission 2018). The general AI sought by AI pioneers (and still being developed today, albeit at a slower pace) would be capable of competing with human intelligence: it would learn and adapt to new situations and different problems. It should, however, still be distinguished from artificial "superintelligence," AI past the point of singularity: the point after which it would greatly surpass human intelligence (Boden 2016).

Contrarily, narrow AI is generally designed to perform limited tasks (e.g., facial recognition or web search) and is increasingly successful in doing so. The task is performed within relatively narrow constraints and parameters. Narrow AI cannot be used for more complex tasks or easily move from one task to a different one: its success is dependent on it remaining in its own specialized area (Boden 2016; Franklin 2014).

All modern tools, including the ones being used and developed for sentencing purposes, fall within this last category. Sentencing machine learning-based models are not generally intelligent: they operate within the preconceived or pre-learned parameters and are unable to adapt to new situations and different problems. In the comparison of human and AI judges, however, we will consider both the existing narrow AI and the developing futuristic general AI, pointing out the differences this distinction involves for our debate.

12. 2 Legitimacy of Sentencing

Legal punishment "is (a) unpleasant, (b) imposed for conduct that has breached legal rules, (c) targeted against the individual responsible for that conduct, (d) imposed intentionally by State agents other than the subject, who are (e) acting under the authority of the breached law" (Hayes 2018, p. 236). Deciding whether to deprive people of their liberty is one of the most difficult decisions we make as a society and for most modern societies, the

most severe restriction of human rights we can imagine. Considering the harm that is inflicted upon the defendant (but not only the defendant!) punishment requires justification; without it, such behavior would be regarded as wrong or evil (De Keijser, Van der Leeden, & Jackson 2002). People have sought justifications for punishment in many different places and ideas. Society (and academics') views on why and how we punish have evolved and are still evolving—with more and more facets being uncovered and alternatives to traditional views being developed. What is often lacking, however, are concepts by which we legitimize the act of sentencing itself. If we accept that punishment is an (essential) part of society, we need to discuss how coming to that punishment needs to be accomplished in order for the punishment and the process to be viewed as legitimate. Therefore, in order to assess whether AI judges are better at achieving legitimacy in sentencing than human judges, we first develop a model in which legitimate sentencing is partitioned into different elements. This model is abstract and a significant simplification of reality, but enables us to analyze the different aspects of legitimacy of sentencing. The first step in the development of the model is distinguishing normative from empirical legitimacy, building upon Roberts and Plesničar (2015).

12.2.1 Normative Legitimacy

Normative legitimacy means that sentencing must have a coherent moral justification. Moral legal theories serve as a critical standard against which sentencing practices are to be judged (De Keijser et al. 2002). The two main moral justifications for criminal punishment are the retributive and the utilitarian. Retribution requires that the severity of the punishment is proportionate to the severity of the crime and the blameworthiness of the offender (von Hirsch 1992). A punishment imposed with utilitarian aims is justified if it maximizes the happiness in society (Michael 1992) taking into account the various costs (e.g., financial, social, emotional to the offender or their family) and benefits (e.g., crimes prevented) of imposing a punishment (Ewald & Uggen 2012). Deterrence, incapacitation, and rehabilitation are utilitarian sentencing strategies. In reality, systems usually have a mixed or hybrid justification model, in which the retributive and utilitarian approaches are combined. Moreover, in recent decades alternative justifications have emerged, including restorative and therapeutic justice—which can hardly be

reconciled with classical justifications for punishment (Snedker 2018; Strang & Braithwaite 2017).

When assessing in more detail whether sentencing practices are normatively legitimate, there are several questions in need of answering. The model we developed to evaluate the normative legitimacy of sentencing distinguishes three different levels: (1) the fundamentals of the system, (2) the actual sentencing decisions as regards the principles, and (3) the effects of the principles in practice.

12.2.1.1 The Fundamentals of the System

The first set of questions addresses the foundation of the system: Is it grounded in moral principles? Are, for example, the aims of sentencing stipulated in the law? Or are there sentencing guidelines that promote moral legal theories for sentencing, for instance, by reflecting ideas about proportionality? Are these ideas and aims coherently implemented throughout the system, and thus providing a coherent framework of sentencing?

The extent to which a sentencing system is grounded in moral theories differs among countries. Systems also differ: some are based on retribution, others have rehabilitation as the core objective, yet others have a mixed set of rationales for punishment. And some systems do not explicitly state the moral theories they are grounded in (Tonry 2011).

The differences between the systems in different countries show that there is no one universal way to ground a sentencing system in moral theories: the moral fundamentals of the system depend on context and culture. However, some sort of moral grounding is vital for both a sense of justice and for a functioning sentencing system. Ashworth (2010), for example, believes that not having a clear sentencing ideology undermines the rule of law, as too much discretion is left to sentencers: not just in terms of adjusting the sentence to the circumstances of the case, but by allowing (too much) space for potential personal beliefs to replace a common rationale. Moreover, not having a clear idea about what sentencing aims to achieve is a cause of disparity (Henham 2013; Hogarth 1971; Palys & Divorski 1986; Wandall 2008).

12.2.1.2 Sentencing Decisions and Principles

Our *second* level of normative legitimacy surrounds the question of the sentencers' attitudes to moral principles: When judges make their sentencing decision, what principles do they apply? And are these the same principles underlying the system?

If judges apply the moral principles that lie at the foundation of the system, normative legitimacy is enhanced. But judges may also intentionally or unintentionally deviate from the system's moral foundations by promoting other sentencing goals (De Keijser et al, 2002; Greenblatt 2008; Morris 1974). Normative legitimacy can still be achieved: what matters is that sentencing decisions reflect moral principles. A lack of explicit moral justification at the foundational level might even be compensated by a strong application of moral theory at the level of the sentencing decision. However, this can also work the other way. A system can have a strong moral foundation that is not discernible in the individual sentencing decisions of the judges. Then there is no normative legitimacy at this second level.

12.2.1.3 Effects of Principles

The *third* question refers to the extent to which the sentencing goals are actually met—or whether they can be met at all: What are good intentions if they do not bring the intended consequences? For a system to be normatively legitimate, the purposes that the system sets out in theory thus need to be met in practice as well; this is the requirement that makes the punishment neither wrongful nor evil (De Keijser et al. 2002). If punishments are meted out with the goal of rehabilitation or deterrence, for instance, have future crimes actually been prevented by the imposition of the punishments?

12.2.2 Empirical Legitimacy

However, as hard as it seems to fulfill our model's elements of normative legitimacy, fulfilling them does not ensure that a sentencing regime is perceived as legitimate. In order to be perceived as such, the sentencing system must also be aligned with the views of the public, a trait we call *empirical legitimacy*. Such alignment with public views will enhance compliance with the law and cooperation with the criminal justice system (Roberts and Plesničar 2015). If people perceive the courts to impose inappropriate sentences or to take into account the wrong factors, the legitimacy of the entire system may be called into question (Henham 2013). In order to have a sentencing regime that is aligned with public views, it needs to be clear and transparent, consistent in application, and sensitive to the input of all relevant parties (Roberts and Plesničar 2015). Effective communication is key. Consequently, we expand our model to assess the legitimacy of sentencing by distinguishing

three requirements for sentencing to be empirically legitimate: (a) transparency, (b) consistency, and (c) communicative effectiveness.

12.2.2.1 Transparency

Legitimate sentencing requires that the public understands why a certain punishment is imposed. In his chapter in this book, Ryberg (2020) explains that clarity and transparency of the sentencing decision are needed to improve the quality of the decisions, to provide accountability of the decisions, and to guide the general public's moral compass as well as manage the public's expectations. They may promote confidence and perceived legitimacy by contributing to a better public understanding of sentencing (Ryberg, 2020).

12.2.2.2 Consistency

The second requirement for legitimate sentencing is that sentencing decisions must be consistent and thus predictable. Similar cases should be punished similarly, and dissimilar cases should be punished dissimilarly to the degree of their dissimilarity. Judges use their discretionary powers to fit the punishment to the case at hand (Saleilles & Ottenhof 2001; Sutherland, Cressey, and Luckenbill 1992). However, this individualization of punishments can undermine legitimacy, if disparity in outcomes cannot be explained by legally relevant factors. For example, when the mood of the judge has affected sentencing, the punishment should not depend on whether the judge suffers from a headache, stress, tiredness, or relationship problems. Disparity in outcomes between judges is also unwarranted: it should not matter if one is sentenced by judge A or judge B. Judges have to be impartial professionals who only take legally relevant factors into account. Decision-making without bias is not only important for the acceptance of the sentencing practices by the public at large (cf. empirical legitimacy) but also for the acceptance of the punishment by the defendant (cf. Tyler's (2003) procedural justice).

12.2.2.3 Effective Communication

The last element of legitimate sentencing is the ability of the system to foster good communication. The public needs to know how and why punishments are meted out and how the sentencing system is performing. In addition, the public also needs to feel able to, within limits, influence how the system is shaped. Good two-way communication is key then. Moreover, effective communication does not only apply to the public, but more importantly, to the participants in the process: the defendant, the prosecution, the victim,

and the witnesses. For a procedure to be considered fair, one of the crucial elements is that people have an opportunity to participate in the situation by explaining their perspective and expressing their views about how problems should be resolved (Tyler, 2003). This is a key element to achieve procedural justice: defendants who perceive that they have been treated with respect and fairness by courts are likely to be more cooperative and compliant with the law and its various agents than those who perceive they have not been treated respectfully and fairly (Walters and Bolger 2019).

12.2.3 The Legitimacy of Sentencing Model

To evaluate whether AI is better at achieving legitimacy in sentencing than human judges, we have thus developed a multilayered model that distinguishes empirical from normative legitimacy (see Figure 12.1). However, albeit separate, these two pillars of our model are interrelated. Public views on sentencing can affect the moral principles that are strived for in the foundation of the sentencing system. For example, public concern over released sex offenders can result in changes in the legal framework that increase the possibilities for incapacitation (McDonald 2012). Public views can also affect the implementation of the moral principles at the level of the actual sentencing decision, for example, when the public demands harsher punishment (Cochran et al 2020). There are three elements of normative legitimacy: (i) the moral principles in the foundations of the system, (ii) the extent to which they are applied through actual decision-making, and (iii) the extent to which sentencing decisions achieve the sentencing goals. These

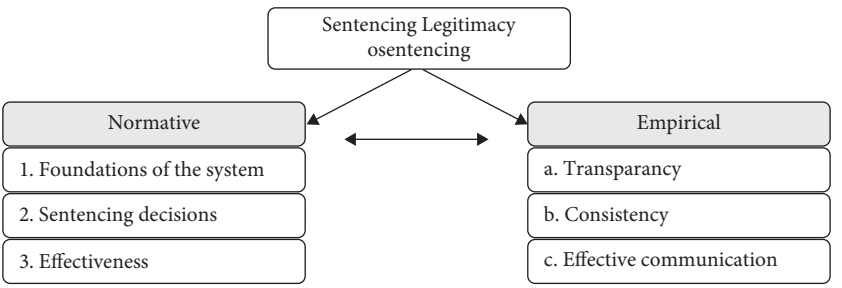

Figure 12.1 Elements of legitimacy at sentencing

all affect public views of sentencing. They are the basis on which public views rest. Normative and empirical legitimacy are thus interrelated.

Generally, the more the individual elements are present in the system, the more legitimate it is and vice versa. We see legitimacy as a continuum, although there are obviously endpoints, where a system might be fully legitimate at one end, and completely illegitimate at the other. Having set this abstract model as the background to our further discussion, we now evaluate whether human or machine judges are better at achieving legitimacy in sentencing.

12.3 Humans vs. Machines

12.3.1 Normative Legitimacy: Foundations of the System

The first level is that of the foundation of the system: Is it coherently grounded in moral principles? We feel that at this level, there is currently no role for AI. Humans have created sentencing systems and chosen the moral theories underpinning these systems. There is currently no realistic prospect of AI assuming those tasks.

But should we allow AI to adjust the foundations of our systems? This requires moral judgment in a novel situation, and at present, AI does not have these capacities (Donohue 2019). However, perhaps a future version of Hal-like super-AI past the point of singularity could develop the ability to make moral judgments and create new conceptions of justice. While this is unlikely for the foreseeable future, we feel very reluctant to accept that this super-AI morality would surpass human morality. Such important decisions about the essence of humanity and the fundamental elements of sentencing should not be left to algorithms that we would not even be able to understand—no matter how well they may perform.

12.3.2 Normative Legitimacy: Sentencing Decisions

The second level of normative legitimacy considers the sentencing decisions of judges: Are humans better at making morally justified decisions than machine judges?

Donohue (2019) asserts that sentencing comes down to a singular moment of moral judgment involving the jurist and the defendant. To make a normative assessment about the punishment to impose, the judge must understand the significance of conceptions of blame, desert, responsibility, excuse, and so forth (Chiao 2019). Virtues such as empathy and compassion are also important for moral judgment, as are intuition and understanding of the human condition (Donohue 2019). Human judges have these capacities although the conceptions differ among judges. Making a moral judgment is inherently subjective. Further, human judgment is responsive to an indefinitely broad range of relevant factors and hence is suited to addressing decision-making contexts in which each case is unique (Chiao 2019). Human judges are also capable of making normative judgments in novel situations (Donohue 2019). Making moral judgments is thus a particularly human ability.

But does the capacity to make moral judgments result in principled sentencing? As de Keijser et al. (2002) have shown, judges may adhere to certain moral principles, but this is not always discernible in their individual sentencing decisions. The punishments judges impose do not always reflect their sentencing goals. Instead of reflecting a consistent moral justification, sentencing seems to be driven by pragmatism and eclecticism (de Keijser et al. 2002). Human judges thus have the capacity to make moral judgments and this allows them to apply moral theories of sentencing when they determine sentence. In practice, however, the sentences they impose are not always in line with these principles.

How about AI? Can AI make moral judgments? Currently, the answer is no. When we consider if/then algorithms, where the rules for sentencing as well as all the normative assessments would have to be programmed into codes, the problems confronting human judges amplify. Legal rules are general and abstract in nature, while algorithms need specific yes/no rules. Moreover, manually programmed algorithms can never fully encompass all the factors and combinations of factors that should affect the sentencing outcome (but see Bagaric and Wolf (2018) for a different view). Not because of the magnitude of these factors and combinations which computers may grasp more efficiently than humans, but because it would require programmers and drafters to have the wisdom to make concise "if/then" rules to convert the moral principles of the system into computer code. Moreover, it is currently impossible to capture human values like empathy and compassion in algorithms. Ethics are too complex to transfer to a computer code (Moor 2006). Moor (2006) notes another difficulty—the computer's absence of

common sense and general knowledge. One could program a machine with Asimov's first law of robotics: "do no harm," but this would only be of use if machines understand the meaning of harm in the real world (Moor 2006).

However, complex AI today works differently and is typically not coded with precise rules. Today's most functional AI is based on the concept of machine learning. At sentencing, this means large datasets of prior judgments are used by the machine to find correlations between characteristics of cases and the imposed punishments on its own—without specific prior "if/then" rules. This learning produces a model able to make decisions and to improve on them by learning from prior experience. This type of algorithmic justice thus tries to replicate prior sentences in similar new cases.

One of the main problems with this approach is that it is based on existing cases. If we wanted the algorithm to propose ideal solutions, these existing cases should have been ideally decided on—but we know this is far from the truth. In fact, as stated before, human sentencing often lacks a moral justification, and many human judgments will probably not be a good basis for principled sentencing. Prior research has shown that these decisions are biased in several ways, for example, with regard to the offender's gender and race (Završnik 2020). These biases will thus be replicated in the decision-making of AI judges. The dangers of "garbage in, garbage out" are apparent here. Moreover, this type of machine learning is less impressive when the past is unlike the future: future cases may be very different from past ones. While humans in these cases make novel decisions, machine learning-based AI will not be able to extrapolate past rules to new circumstances—it is made for replication and not innovation (Fagan and Levmore 2019). This type of AI is not equipped to deal with developments in society or changing views of crime seriousness. As such, it cannot be used for some form of dynamic justice that tracks the developments in the norms of society.

Finally, in the future, general AI should be able to function as an ethical agent. A full ethical agent has human-like qualities: it can make independent moral judgments but also exhibits qualities like intentionality, consciousness, free will, and empathy (Segun 2020). It can also be held accountable for its decisions. The question is then whether virtues such as empathy may also be self-learnable for general AI. And if so, should humans then accept the punishments handed out by full ethical agent machine judges as superior to human decision-making? Again, our strongest objection against the idea of AI making sentencing decisions is that this function is too important to be entrusted to an entity whose reasoning we do not understand.

12.3.3 Normative Legitimacy: Effects of Principles

After the foundation of the system and the actual sentencing decisions of judges, the final level in our model of normative legitimacy is the effect level: Are the moral principles of the sentencing system met in practice? Are the sentencing goals that judges aimed for achieved?

In theory, a system founded in retribution, or judges aiming for retribution, is effective when sentencing is proportionate to the seriousness of the crime and the blameworthiness of the offender. But in practice, this is difficult to assess: What is a proportional punishment? What is considered to be a just punishment is ambiguous and not universal: it is culturally dependent and subjective (Plesničar & Šugman Stubbs 2018). An evaluation of the severity of the harm that the offender caused by his crime requires an evaluation for which subjective sentiments like empathy and compassion are key. There is no objective answer to the question of what a proportionate punishment exactly is, beyond the sentencing ranges set out for individual offenses. While the concepts of cardinal and ordinal proportionality help to promote proportional sentencing, we cannot say the same about individual sentencing decisions. We thus find it hard to evaluate the extent to which human judges are capable of meting out proportionate punishments. But since imposing a proportionate punishment requires moral considerations, we think it safe to assume that machine judges (who lack the ability for moral judgment) will perform worse than humans.

If we look at the utilitarian perspective, effects might be easier to assess—in theory. In theory, a punishment is justified from a utilitarian perspective when it maximizes happiness in society. In practice, this justification is hard to assess, since the costs and benefits are unknown, or even unknowable to the judge: How many crimes are prevented if the offender is locked up for a certain amount of time? And how much weight should be given to, for example, a prevented rape? Or, moreover, how much weight should be assigned to the suffering of the defendant's children left without a parent for a prolonged time? If the costs and the benefits are unclear, how can we tell whether the punishment was effective? Moreover, recidivism rates are high, so the effectiveness of punishment in terms of deterrence and rehabilitation is questionable. And if punishments imposed with the aim of preventing future crimes are not effective, their imposition cannot be morally justified at this level of the model. With the current knowledge about the consequences of punishment, achieving legitimacy at this effect level of the model seems almost impossible, both for human and AI judges.

Perhaps this is where future AI may prove most effective. By connecting many different datasets that contain information on the consequences of punishment, AI could develop more insights into the effects of punishment. AI could, for example, follow offenders over time to determine whether they commit new crimes, whether they have a job, where they live and—by connecting to social media—even see what their social networks look like and what their hobbies are. This information on the offender's progress could be crucial to learning about the effects of punishment. Machine judges could use this information to learn about which offenders are deterred and when. Or to learn about the effectiveness of different rehabilitation programs for different types of offenders. Connecting all these datasets and letting AI use it for sentencing might increase the effectiveness of sentencing, but comes with high social costs, like privacy issues. Again, the issue is this: Do we want to entrust AI with this power?

12.3.4 Empirical Legitimacy: Transparency

As noted earlier, clarity and transparency of sentencing decisions are both important for legitimacy in sentencing. Clarity is needed to check how moral justification theories are applied in the actual sentencing decisions, while transparency allows the public to see how decisions are made.

Human judges provide insight into their sentencing decisions by giving reasons. However, these reasons do not explain everything. The same rationale, even the exact same wording, can be used to justify a prison sentence of six, seven, or eight years. Moreover, the judge can only refer to characteristics of the crime or of the offender that she *consciously* took into account. However, we know that sentencing decisions are also shaped by factors that the judge unconsciously considers, for example, because of stereotyping (Albonetti 1991). When the judge is unaware of the factors influencing her decision, she cannot account for them (see the contribution of Chiao (2020) to this book for examples of unconscious influences on sentencing decisions). Thinking thus remains hard to explain and giving reasons is not the same as reasoning. There is "no generally accepted theory of how cognitive phenomena arise from computations in cortex" (Valiant 2014, 15). What happens in the mind of the judge when she makes her sentencing decision is a black box in itself.

Moreover, at the systemic level, modern sentencing systems are complex and difficult to comprehend. The problem goes beyond just judges having

difficulty explaining their decision-making. For a layperson, understanding how sentencing operates is often hard: accounting for different levels of culpability, combinations of factors, etc., makes for a very complex system, one which is far from clear or transparent.

The two issues combine when stepping into the zone of AI judges. Individual decisions should be the direct result of what the system had predicted. The transparency of the algorithms of machine judges depends on the type of AI that is used. "If/then" algorithms (that are not yet AI) producing decision trees are very transparent, although extensive codes can make it difficult to see the forest for the trees. Models based on machine learning are much less transparent, especially because it is not always or immediately clear *why* certain factors have a certain value for the sentencing decision. Machine learning looks for relations between characteristics that best predict the outcome. It does not look for characteristics that should affect the punishment according to legal principles. The transparency is limited to showing which factors predict the sentencing outcome, and this makes the sentencing decision unclear. Moreover, today's most effective machine learning uses the black box method: data come into the model and results come out—the process and method by which that happens are not the focus of the task. Further developments might improve this current lack of transparency, but for now, it seems an illusion (Goebel et al. 2018).

Clarity and transparency would be even more of an issue with futuristic AI. Such AI might then be the only one who understands how the system works: not even experts in the field of computer science could audit the process. In that case, no one could assess whether the sentencing determinants are reliable at all (Chiao 2019). This is the type of opacity that Ryberg calls "technically caused opacity" in his chapter.

Transparency is thus not well achieved by human judges. For AI judges, while simplistic algorithms can be very transparent, more sophisticated or self-learning AI will necessarily lead to further opacity.

12.3.5 Empirical Legitimacy: Consistency

Sentencing should not only be transparent, but it should also be consistent and thus predictable. Similar cases should be punished similarly, and only legally relevant factors should be taken into account by the judge. For human judges, consistency in sentencing is a challenge because human reasoning

is flawed. Setting aside intentional wrongdoing, human judges are as prone to making judgment mistakes as humans in general (Schauer 2010). For example, cognitive biases may very likely affect judges' decision-making when assessing the blameworthiness and dangerousness of the offender. Such decisions are made in a context where time and information are limited (Albonetti 1991; Steffensmeier, Ulmer, and Kramer 1998). In order to deal with these uncertainties, judges develop a decision-making schema that draws upon past experiences and societal stereotypes to determine the defendant's risk and blameworthiness. Relying on stereotypes could be one of the causes of unwarranted sentencing disparity. Research has shown for example that—in some contexts—Black defendants are punished more severely than White defendants, and male defendants more severely than female (Baumer 2013; Bontrager, Barrick, and Stupi 2013). Research has also demonstrated differences in sentencing outcomes between judges in similar cases (Spohn 2008; Wooldredge 2010). Subjective assessments of the facts of the case, the relevant circumstances of the offender, the preferred sentencing goals, and the punishment that is deemed just may cause sentencing disparity.

Disparity was the main reason for the introduction of sentencing guidelines across common law jurisdictions (Ashworth 2009; Frankel 1972; Stith & O'Neill 2003). Sentencing guidelines limit the discretionary powers of judges to leave less room for a subjective assessment to affect sentencing outcomes. This is particularly true for the restrictive sentencing grids that employ only two dimensions: crime seriousness and criminal history. Once these are determined, a court must impose a sentence within a specific range, or provide compelling reasons to impose a different sentence (Frase 1990). In these systems, instead of executing moral judgment, judges became discretion-less "accountants" in a scheme set up by others (Donohue 2019). Judges objected to the guidelines because they could not individualize sentences, they had to impose punishment that did not feel just (Leipold 2005; Stith and Cabranes, 1998). Consistency was not to be reached at the expense of discretion.

Inconsistency in sentencing by human judges was one of the main reasons to employ computers to decide upon the sentence. AI judges are not prone to cognitive biases or subjective assessments. They are never tired, hungry, cranky, bored, or stressed (cf. Danziger, Levav, and Avnaim-Pesso 2011). A negative emotional state of mind—or a positive one—cannot affect sentencing outcomes, thereby reducing the disparity in sentencing.

But algorithms designed by humans seem more objective than they really are. Who makes the decision rules for the algorithm? Disparities among ethical theories make it a daunting task to embed ethics or a moral code into AI systems (Segun 2020). And aren't these rules likely to reflect the biases of those who develop them (Estcourt and Marr 2019)? And what are the risks of having legal rules and normative considerations being translated into computer language by a programmer who knows nothing about the law?

Machine learning AI is also less objective than it seems because the models are built on data infected with bias. Machine learning replicates these existing biases: biases present in the verdicts of the judges become embedded in the algorithm. These algorithmic biases affect sentencing outcomes and reinforce existing inequality and stereotypes (Segun 2020). Ironically, machine learning AI ends up worsening disparity instead of reducing it. This is not due to a mechanical flaw; human judges are to blame. There is no database from which the algorithm could learn that would be free of existing biases. It is, however, a problem that machines cannot solve for humans.

One solution to this problem of algorithmic bias would be to program the decision-making AI in a way that would neutralize biases (Chiao 2019). This would only be possible if the algorithms were transparent and comprehensible. In consultation with judges, experts could program the machine to disregard extralegal sentencing factors such as race, ethnicity, and gender, even if they enter through data learning. However, there is more than just a technological problem here. Not only would such AI be at odds with the most efficient models of modern AI (which include deep learning), but there is a deeper problem: it is unclear what just punishment requires. If racial disparity is to be eliminated, should Blacks receive a discount? Or Whites a severity premium? Using machine judges to neutralize existing biases is thus doubly problematic. First because of the complexity of the algorithms—a problem that may be overcome in the future. Second, because it requires the principles that the system is grounded in to be translated into computer code.

12.3.6 Empirical Legitimacy: Effective Communication

The final element of empirical legitimacy is effective communication. Communication is what has evolutionarily made us thrive (Harari 2017).

Communication at sentencing, however, is a more complex issue. The roles of various participants at sentencing have been evolving. Victims, for example, once completely excluded (Christie 1977), now enjoy participatory rights through victim statements (Roberts and Erez 2004) or the right to appeal (Briški 2020). The various participants that judges need to include in the communication and their interrelationships make it hard for judges to determine how and when to include them. This may leave participants disappointed and feeling "unheard," which in turn undermines perceptions of legitimacy. However, human judges are generally well equipped to address various participants as fellow moral agents. With regard to the defendant, for example, they can express censure in the expectation that the defendant will understand and internalize the message. This is an essentially human activity (Chiao 2019), even if some defendants seem insensitive to the moral message.

For AI judges, communication may be their biggest obstacle. There is no real communication apart from imputing the relevant data. There is no listening or "being heard," there is no empathy or sympathy from the judge—for example when witnesses are testifying about traumatic events. People react to how they feel and experience events and settings. Doctors are, for example, much less likely to be sued for malpractice when they evince empathy toward victimized patients (Smith et al. 2016).

The criminal process and sentencing in particular, have far-reaching consequences (Tata 2020). That process—of one robed judge and one convicted defendant in conversation—has moral value in and of itself, and the addition of an interloping machine may undermine that function (Donohue 2019).

When considering machine judges' decisions and their ability to persuade the participants of the procedure or the public, the concept of algorithm aversion seems an important one to consider (Burton, Stein, and Jensen 2020; Dietvorst, Simmons, and Massey 2015). Algorithmic aversion is a bias that causes humans to mistrust algorithmic decisions simply because they are not human—despite AI's record of sounder decisions. When considering AI-based decisions, the margin of error humans are willing to tolerate is none. In criminal justice, the lack of human interaction and the "dehumanizing" effect this could bring be an insurmountable problem (for a different view, see Bagaric, Hunter, and Stobbs (2019)).

12.4 Conclusion

Would you rather be sentenced by a human or a machine? We based our comparison between human and AI judges on an abstract model of legitimacy and find AI rather lacking on several levels. First, we argued that current AI is incapable of making moral decisions and this is crucial for our concept of normative legitimacy. Legitimacy at the level of the foundation of the system is thus best achieved by humans. Futuristic AI might eventually develop into a full moral agent. Still, even then, we feel that decisions about the moral foundation of the sentencing systems should not be entrusted to AI. At the second level, that of judges' actual sentencing decisions, we reasoned that in order to make a principled sentencing decision, the judge needs to apply moral principles. And again, we argue that current AI does not understand morality.

Our final assessment criterion for normative legitimacy concerns effectiveness: Were the goals achieved? We conclude that it is difficult to assess whether judges succeed in imposing proportionate punishments since it is unclear how severe a punishment must be in order to be proportionate to the seriousness of the crime and the blameworthiness of the offender. And the effectiveness of utilitarian sentencing can only be assessed if the consequences of the punishment for the offender and society at large are known. So for both human and AI judges, the extent to which their decisions achieve the desired effects is unknown. However, future AI using "big data" might gain insight into the consequences of punishment. Futuristic AI is then probably best able to achieve utilitarian sentencing goals—but at great cost in terms of our privacy.

Regarding empirical legitimacy, we conclude that human judges cannot fully explain their reasoning. A judge is as much of a black box as a complex AI model. Conversely, simple algorithms used by machine judges can be very transparent. But as the algorithms become more complicated, their decisions become more opaque. This is especially true for self-learning AI.

We conclude that machine judges are better at achieving consistency than humans. But self-learning AI suffers from *algorithmic bias*: inconsistencies that were already prevalent in the data (previous sentencing decisions) are replicated and reinforced. When this occurs, sentencing decisions become consistent but consistently wrong. Futuristic general AI might be able to recognize and correct the bias in sentencing, but the codes of the machine would be so complex that understanding why certain (combinations of) case

characteristics result in sentencing discounts or premiums would be impossible. Improving this level of empirical legitimacy would thus necessarily lessen transparency.

The third element of empirical legitimacy concerns effective communication. Here we conclude that humans surpass AI judges. Human judges are able to communicate effectively with participants and the public and engage with them as one moral agent to another. While some human judges are better at it that than others, AI judges may never achieve it at all—humans do not perceive them as equal decision-makers.

In conclusion, we regard AI judges as incapable of generating normative legitimacy. Even when this might be achievable in the future, we have serious reservations about letting it establish a new moral philosophy for sentencing. Either this comes at a significant cost (such as at the level of effects) or entrusts too much power to mechanisms we do not fully understand and thus cannot thoroughly assess. Moreover, it might be too easy for humans to shift such responsibility to AI. Relieving humans from serious consideration of the morality of punishment would allow us to inflict pain (legitimate and legal, but still pain) without feeling in any way responsible (cf. Floridi et al. 2018).

However, at the level of empirical legitimacy, the use of AI may improve matters. While transparency is still lacking, further development might bring improved results in that area. Moreover, consistency is AI's strength—and while today's systems are flawed due to corrupt learning datasets, further development might improve that. Transparency and consistency are major challenges for human judges as well. But effective communication, conversely, is where AI judges fail. Even if they were able to learn empathy and even compassion in making their sentencing decisions, they would not be able to convey it in an effective and approachable manner.

Huq (2020 639) raises another important question relevant to the choice between humans and machines: Are AI's flaws easier to identify and remedy than the flaws of its human analog? Translated to our context: Is it easier to teach a machine how to make principled decisions than to teach a human judge to sentence consistently? It far from easy to do the latter. And while it is extremely hard to do the former today, it might become easier over time. Should we then leave open the option of handling sentencing to AI in the future?

Estcourt and Marr (2019, 856) think that while many decisions could be handed over to machines, "only some of them should be, even when the

machines can make them better than we do." This seems contradictory—why do something badly when you have the option of doing it better? We think, however, that it is not just a matter of good or bad sentencing outcomes. There are crucial features of sentencing that are so inherently human that we cannot imagine them being successfully replaced by AI. Making moral judgments not only requires us to consider various justifications for punishing people but also makes us take responsibility for our actions. If we leave sentencing to AI—are we not losing the very essence of what makes sentencing a human process?

References

Albonetti, C. A. 1991. "An Integration of Theories to Explain Judicial Discretion." Social Problems, 38 (2): pp. 247–266.

Ashworth, A. 2010. Sentencing and Criminal Justice. Cambridge: Cambridge University Press.

Ashworth, A. 2009. "Techniques for Reducing Sentence Disparity." In Principled Sentencing: Readings on Theory and Policy, edited by Andrew Von Hirsch, Andrew Ashworth, and Julian V. Roberts, pp. 243–258. Oxford: Hart.

Bagaric, M., and G. Wolf (2018). Sentencing by Computer: Enhancing Sentencing Transparency and Predictability and (Possibly) "Bridging the Gap between Sentencing Knowledge and Practice." George Mason Law Review 25: pp. 653–710.

Bagaric, M., D. Hunter, and N. Stobbs. 2019. "Erasing the Bias Against Using Artificial Intelligence to Predict Future Criminality: Algorithms Are Color Blind and Never Tire." University of Cincinnati Law Review, 88 (4): pp. 1037–1081.

Baumer, E. P. 2013. "Reassessing and Redirecting Research on Race and Sentencing." Justice Quarterly 30 (2): pp. 231–261.

Boden, M. A. 2016. AI: Its Nature and Future. Oxford: Oxford University Press.

Bontrager, S., K. Barrick, and E. Stupi. 2013. "Gender and Sentencing: A Meta-Analysis of Contemporary Research." Journal of Gender, Race & Justice 16: pp. 349–372.

Briški, L. 2020. "Oškodovančev vpliv na odločitev kazenskega sodišča v slovenskem in nemškem kazenskem postopku [The victim's influence on the decision of the criminal court in Slovenian and German criminal proceedings]." Pravna praksa 39: pp. 8–9.

Burton, J. W., M. K. Stein, and T. B. Jensen. 2020. "A Systematic Review of Algorithm Aversion in Augmented Decision Making." Journal of Behavioral Decision Making 33 (2): pp. 220–239.

Chiao, V. 2019. "Fairness, Accountability and Transparency: Notes on Algorithmic Decision-Making in Criminal Justice." International Journal of Law in Context 15 (2): pp. 126–139.

Chiao, V. 2020. "Transparency: Are Judges Better Than Algorithms?" In Principled Sentencing and Artificial Intelligence, edited by Jesper Ryberg and Julian V. Roberts, pp. 34–57. Oxford: Oxford University Press.

Cochran, J. C., E. L. Toman, R. T. Shields, and D. P. Mears. 2020. "A Uniquely Punitive Turn? Sex Offenders and the Persistence of Punitive Sanctioning." Journal of Research in Crime and Delinquency. July: pp. 1–45. doi:10.1177/0022427820941172

Christie, N. 1977. "Conflicts as Property." The British Journal of Criminology 17 (1): pp. 1–15.

Danziger, S., J. Levav, and L. Avnaim-Pesso. 2011. "Extraneous Factors in Judicial Decisions." Proceedings of the National Academy of Sciences 108 (17): pp. 6889–6892.

De Keijser, J. W., R. Van der Leeden, and J. L. Jackson. 2002. "From Moral Theory to Penal Attitudes and Back: A Theoretically Integrated Modeling Approach." Behavioral Sciences & the Law 20 (4): pp. 317–335.

Dietvorst, B. J., J. P. Simmons, and C. Massey. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err." Journal of Experimental Psychology: General 144 (1): p. 114.

Donohue, M. E. 2019. "A Replacement for Justitia's Scales? Machine Learning's Role in Sentencing." Harvard Journal of Law & Technology 32 (2): pp. 657–678.

Estcourt, A., and K. Marr. 2019. "Thinking Machines and Smiley Faces." Australian Law Journal 93 (10): pp. 855–865.

EU-Commission. 2018. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions on Artificial Intelligence for Europe, COM/2018/237 final.

Ewald, A., and C. Uggen. 2012. "The Collateral Effects of Imprisonment on Prisoners, Their Families, and Communities." In The Oxford Handbook on Sentencing and Corrections, edited by Joan Petersilia and Kevin R. Reitz, pp. 83–103. New York: Oxford University Press.

Fagan, F., and S. Levmore. 2019. "The Impact of Artificial Intelligence on Rules, Standards, and Judicial Discretion." Southern California Law Review 93 (1): pp. 1–35.

Floridi, L., J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, . . . F. Rossi. 2018. "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." Minds and Machines 28 (4): pp. 689–707.

Frankel, M. E. 1972. "Lawlessness in Sentencing." University of Cincinatti Law Review 41: pp. 1–54.

Franklin, S. 2014. "History, Motivations, and Core Themes." In The Cambridge Handbook of Artificial Intelligence, edited by Keith Frankish and William M. Ramsey, pp. 15–33. Cambridge: Cambridge University Press.

Frase, R. S. 1990. "Sentencing Reform in Minnesota, Ten Years After: Reflections on Dale G. Parent's Structuring Criminal Sentences: The Evolution of Minnesota's Sentencing Guidelines." Minnesota Law Review 75 (3): pp. 727–754.

Goebel, R., A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, . . . A. Holzinger. 2018. Explainable AI: The New 42? Paper presented at the International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Hamburg.

Greenblatt, N. 2008. "How Mandatory Are Mandatory Minimums? How Judges Can Avoid Imposing Mandatory Minimum Sentences." American Journal of Criminal Law 36 (1): pp. 1–38.

Harari, Y. N. 2017. Homo Deus: A Brief History of Tomorrow. New York: Harper.

Hayes, D. 2018. "Proximity, Pain, and State Punishment." Punishment & Society 20 (2): pp. 235–254.

Henham, R. 2013. Sentencing and the Legitimacy of Trial Justice. New York: Routledge.

Hogarth, J. 1971. Sentencing as a Human Process. Toronto: University of Toronto Press.

Huq, A. Z. 2020. "A Right to a Human Decision." Virginia Law Review, 106 (3): pp. 611–688.

Leipold, A. D. 2005. "Why Are Federal Judges So Acquittal Prone." Washington University Law Quarterly, 83: pp. 151–227.

McDonald, D. 2012. "Ungovernable Monsters: Law, Paedophilia, Crisis." Griffith Law Review 21(3): pp. 585–608.

Michael, M. A. 1992. "Utilitarianism and Retributivism: What's the Difference?" American Philosophical Quarterly 29 (2): pp. 173–182.

Moor, J. H. 2006. "The Nature, Importance, and Difficulty of Machine Ethics." IEEE Intelligent Systems 21 (4): pp. 18–21.

Morris, N. 1974. The Future of Imprisonment. Chicago: University of Chicago Press.

Palys, T. S., and S. Divorski. 1986. "Explaining Sentence Disparity." Canadian Journal of Criminology 28 (4): pp. 347–362.

Plant, R. 1994. An Introduction to Artificial Intelligence. Paper presented at the 32nd Aerospace Sciences Meeting and Exhibit, AIAA 1994-294.

Plesničar, M. M., and K. Šugman Stubbs. 2018. "Subjectivity, Algorithms and the Courtroom." In Big Data, Crime and Social Control, edited by Aleš Završnik, pp. 154–176. London: Routledge.

Roberts, J. V., and E. Erez. 2004. "Communication in Sentencing: Exploring the Expressive Function of Victim Impact Statements." International Review of Victimology 10 (3): pp. 223–244.

Roberts, J. V., and M. M. Plesničar. 2015. "Sentencing, Legitimacy, and Public Opinion." In Trust and Legitimacy in Criminal Justice: European Perspectives, edited by Gorazd Meško and Justice Tankebe, pp. 33–51. Cham: Springer International Publishing.

Ryberg, J. 2020. "Sentencing and Algorithmic Transparency." In Sentencing and Artificial Intelligence, edited by Jesper Ryberg and Julian V. Roberts, pp. 13–34. New York: Oxford University Press.

Saleilles, R., and R. Ottenhof R. 2001. L'individualisation de la peine de Saleilles à aujourd'hui; suivie de L'individualisation de la peine: cent ans après Saleilles. Ramonville Saint-Agne: Érès.

Schauer, F. 2010. "Is There a Psychology of Judging?" In The Psychology of Judicial Decision Making, edited by David E. Klein and Gregory Mitchell, pp. 103–120. Oxford: Oxford University Press.

Segun, S. T. 2020. From Machine Ethics to Computational Ethics. AI & Society. Retrieved from https://doi.org/10.1007/s00146-020-01010-1.

Smith, D. D., J. Kellar, E. L. Walters, E. T. Reibling, T. Phan, and S. M. Green. 2016. "Does Emergency Physician Empathy Reduce Thoughts of Litigation? A Randomised Trial." Emergency Medicine Journal 33 (8): pp. 548–552.

Snedker, K. A. 2018. Therapeutic Justice: Crime, Treatment Courts and Mental Illness. London: Palgrave Macmillan.

Spohn, C. 2008. How Do Judges Decide?: The Search for Fairness and Justice in Punishment. Los Angeles: Sage Publications.

Steffensmeier, D., J. Ulmer, and J. Kramer. 1998. "The Interaction of Race, Gender, and Age in Criminal Sentencing: The Punishment Cost of Being Young, Black, and Male." Criminology 36 (4): pp. 763–797.

Stith, K., and J. A. Cabranes. 1998. Fear of Judging: Sentencing Guidelines in the Federal Courts. Chicago: University of Chicago Press.

Stith, K., and M. E. O'Neill. 2003. "Federal Sentencing Guidelines Symposium Yale Law School." November 8, 2002. Federal Sentencing Reporter 15 (3): pp. 156–159.

Strang, H., and J. Braithwaite. Restorative Justice: Philosophy to Practice. London: Routledge.

Sutherland, E. H., D. R. Cressey, and D. F. Luckenbill.1992. Principles of Criminology. Lanham, MD: General Hall.

Tata, C. 2020. Sentencing: A Social Process. London: Palgrave Pivot.

Tonry, M. 2011. "Introduction: Thinking about Punishment." In Why Punish? How Much? A Reader on Punishment, edited by Michael Tonry, pp. 3–28. Oxford: Oxford University Press.

Tyler, T. R. 2003. "Procedural Justice, Legitimacy, and the Effective Rule of Law." Crime and Justice: A Review of Research 30: pp. 283–357.

Valiant, L. G. 2014. "What Must a Global Theory of Cortex Explain?" Current Opinion in Neurobiology 25: pp. 15–19.

Von Hirsch, A. 1992. "Proportionality in the Philosophy of Punishment." Crime and Justice 16: pp. 55–98.

Walters, G. D., and P. C. Bolger 2019. "Procedural Justice Perceptions, Legitimacy Beliefs, and Compliance with the Law: A Meta-Analysis." Journal of Experimental Criminology 15 (3): pp. 341–372.

Wandall, R. H. 2008. Decisions to Imprison: Court Decision-Making Inside and Outside the Law. Aldershot: Ashgate.

Wooldredge, J. 2010. "Judges' Unequal Contributions to Extralegal Disparities in Imprisonment." Criminology 48 (2): pp. 539–567.

Završnik, A. 2020. "Criminal Justice, Artificial Intelligence Systems, and Human Rights." ERA Forum 20 (4), pp. 567–583.