The Right to be an Exception in Data-Driven Decision-Making

Sarah H. Cen
SHCEN@MIT.EDU
MIT

April 12, 2022

Abstract

Data-driven assessments estimate a target—such as the likelihood an individual will recidivate or commit welfare fraud—by pattern matching against historical data. There are, however, limitations to pattern matching. Even algorithms that boasts near-perfect performance *on average* can produce assessments that perform poorly on *specific* individuals. From the assessment's point of view, these individuals are anomalies or *exceptions*, and in some contexts, these failures can lead to decisions that inflict irreparable harm on individuals through no fault of their own. In this Article, we study how overlooking exceptions can yield undesirable outcomes and how this observation already motivates notions in the law—such as dignity and the right to individualized sentencing—as well as research areas in computer science—such as causal inference and robust optimization. Although the belief that exceptions matter to high-stakes decisions is not new, the absence of a legal framework that acknowledges the unique challenges around exceptions in data-driven contexts has left a large accountability gap in the governance of data-driven decisions. To close this gap, this Article proposes that individuals have the *right to be an exception* in data-driven decision-making. The right requires that, when a decision can inflict harm on an individual, the decision maker must consider the level of uncertainty that accompanies a data-driven assessment and, in particular, whether it is appropriately individualized. The greater the risk of harm, the more serious the consideration. In this Article, we unpack the right to be an exception in detail, examining how it necessitates that uncertainty be meaningfully incorporated into data-driven decisions, affects the legitimacy of and trust in algorithms, rebalances the burden of proof between decision makers and subjects, and more. We conclude by discussing ex ante and ex post legal measures and surveying related areas in algorithm design.

1 Introduction

An *exception* is an instance that does not follow a general rule. Although exceptions are rare by definition, they are critical to many *data-driven decisions*, which we define as decisions informed or made by data-driven algorithms.

Consider autonomous vehicles. One of the expected benefits of autonomous driving is that it will reduce fatalities. However, because datasets provided by human drivers rarely contain accidents, especially fatal ones, a driving algorithm that is trained on only day-to-day data may achieve near-perfect performance on average but fail catastrophically when it encounters an accident precisely because accidents are uncommon (Schwarting et al., 2018). As another example, suppose an admitted hospital patient is statistically similar to previous patients. One approach to medical care would proceed by testing and treating the patient as if they embody the average of previous similar patients (Obermeyer et al., 2019). Alternatively, one could rule out that the patient is not an exceptional case, testing and treating in accordance. While operationally similar for most patients, these are distinct approaches to decision-making, and their outcomes differ on the patients who, through no fault of their own, are exceptions.

When machine learning (ML) fails to recognize and treat exceptions appropriately, there is currently no clear legal recourse for the decision subjects. Because statistical averages are so fundamental to modern ML, the absence of a clear legal framework that intervenes when data-driven decisions inflict significant harm due to a failure to account for exceptions has left a large accountability gap in the governance of data-driven decisions. As algorithm designers continue working toward recognizing and treating exceptions appropriately, defining the rights of data-driven decision subjects is a necessary legal complement to these efforts.

In this Article, we examine data-driven decision contexts in which exceptions matter. We study how failing to account for exceptions in these contexts can yield undesirable outcomes and how this observation already motivates

multiple notions in law—including recognition dignity and the right to individual sentencing—as well as several research areas in computer science—including causal inference, robust optimization, and individual fairness. We argue that, when data-driven decisions directly affect individuals and the stakes are high, decision subjects should be afforded the *right to be an exception*.

The right to be an exception does not imply that every individual *is* an exception but that, when a decision may inflict harm on the decision subject, the decision maker should consider the possibility that the subject *may* be an exception. The right to be an exception involves three ingredients: *harm*, *individualization*, and *uncertainty*. The decision maker must choose to inflict harm only when they have considered whether the decision is appropriately individualized and, crucially, the uncertainty that accompanies the decision's data-driven component. The greater the risk of harm, the more serious the consideration. The right to be an exception does not diminish attention to other concerns, such as the cost individualization imposes on the decision maker or the effect of a decision on others. Rather, it requires a level of consideration fitting to the risk of harm, and such consideration can and should account for relevant factors.

Establishing the right to be an exception fills a gap in the governance of data-driven decisions. Legal frameworks designed for human decision makers are ill equipped to handle data-driven decisions because, among other reasons, algorithms can be deployed on large scales and without much clarity into the algorithm's assessment. We show in this Article that a failure to uphold an decision subject's right to be an exception can cause irreparable harm and that this failure is not fully addressed by legal systems designed for human decision makers. To close this gap, the right to be an exception provides a basis for individuals to contest such decisions. By shifting power away from decision makers who are currently permitted to argue that excellent (or even good) average-case performance justifies the poor treatment of the few exceptions, it also rebalances the burden of proof. Beyond its benefits to decision subjects, establishing a right to be an exception can also benefit society at large by enhancing the legitimacy of algorithms that uphold this right, improving trust in these algorithms, and ultimately encouraging better algorithm design.

This Article is organized as follows. In Section 2, we position the right to be an exception relative to technical and legal concepts. In Section 3, we lay out the right to be an exception in detail, outline criteria for when it should be applied, and explain its three main components—harm, individualization, and uncertainty. We also examine the potential benefits of the right to be an exception, touching on issues of contestation, trust, legitimacy, burden of proof, and transparency. We conclude in Section 4 with a discussion on how the right to be an exception could be operationalized, including ex ante and ex post legal measures as well as directions in algorithm design. We note that this Article is intentionally cross-disciplinary and blends analyses from both computer science and law.

2 Background

In this section, we examine how exceptions arise in data-driven decisions from a technical perspective, then shift to the basis for the right to be an exception from a legal perspective. We find that the concept of *dignity* is closely related to the right to be an exception and conclude with an example of another right that invokes dignity.

2.1 Why exceptions arise in data-driven decision-making

At a high level, ML seeks to learn a model f that, when applied to inputs $\mathbf{x}$, produces an output $f(\mathbf{x})$ that is "close" to the true target (Bishop, 2006). For instance, given meteorological data $\mathbf{x}$ that is gathered on Sunday, f produces a prediction $f(\mathbf{x})$ of the chance of rain on Monday.

The model f is chosen from a class or family $\mathcal{F}$ of possible models. Among the possibilities in $\mathcal{F}$, the chosen model f is that which maximizes or minimizes a predetermined objective function R. For example, one may choose the weather model that most accurately predicts the chance of rain according to historical data $\mathcal{D}$, in which case $R(f)$ measures prediction accuracy of f over data $\mathcal{D}$, and the chosen model f achieves the maximum accuracy over all possible models $\mathcal{F}$. Alternatively, one may decide to choose the weather model that maximizes the happiness of the weather forecast's readers based on historical data $\mathcal{D}$. This objective is usually distinct from accuracy (cf. scoring rules (Parmigiani and Inoue, 2009)) because the level of unhappiness that people experience when caught in a downpour without an umbrella (one type of error) is different from the level of unhappiness they feel when they have no need for an umbrella that they packed (another type of error). In this case, the objective $R(f)$ measures the aggregate happiness that a model f would have induced over historical outcomes $\mathcal{D}$. Although this framework–consisting of five main elements: input, output, model class, objective function, and data—is fairly simple, it can produce a wide range of possibilities, as exhibited by the breadth of ML.

This Article is motivated by the observation that *modern ML is built on averages.* The most popular methodologies—including empirical risk minimization (Devroye et al., 1997), maximum likelihood estimation (Devroye et al., 1997), and regret minimization (Sutton and Barto, 2018)—are centered around average-based objective functions. Moreover, the vast majority of metrics used to evaluate models—including accuracy, precision, recall, and success rate (Fawcett, 2006; Silver et al., 2017)—are all average notions. There are many reasons averages are so popular, one of which is the foundational belief that, with sufficiently rich inputs, models, and data, the model that performs best on average will produce perfect assessments. Intuitively, this makes sense. With enough information and computation, one can use averages to make reasonable judgments about most cases. In the limit of infinite information and computation, averages are taken over increasingly fine-grained inputs (e.g., weather conditions), each with enough historical data on what they imply, that one should be able to make perfect judgments for not only most but all cases. However, as discussed in Section 3, this holy grail outcome is not always achievable, especially when the stakes are high. The cases on which ML methods fail are what we call *exceptions.*

There are multiple ways exceptions arise from a technical standpoint, and we briefly describe four below.

1. *Sampling bias* causes one type of sample T to appear fewer times in the data than other types (Cuddeback et al., 2004). One common instance of sampling bias are outliers: events that occur low probability in a dataset. For instance, suppose that a medical condition T is rare and appears only once in a dataset collected over the general population. Unless an algorithm accounts for the fact that T is rare in this dataset, training on this dataset using an average-based approach can result in poor performance for patients of type T.

2. *Model capacity* is a measure of a model's expressiveness (Vapnik et al., 1994; Hu et al., 2021). Exceptions can occur when a model's capacity is too low to capture patterns beyond broad-strokes generalizations. As an example, a neural network's capacity is determined by the number of nodes and layers. If the relationship between the input variables $\mathbf{x}$ and the target variable y is more complex than the expressiveness of the model, then the model must make simplifications. For instance, if the relationship between $\mathbf{x}$ and y is quadratic, then a linear function would not be able to capture the relationship between $\mathbf{x}$ and y in its entirety. At best, f may capture the approximate relationship between $\mathbf{x}$ and y for a range of $\mathbf{x}$-values but not all. The model can therefore perform well on some inputs but at the cost of performing poorly on others, and exceptions are cases $\mathbf{x}$ for which the model does not perform well.

3. *Distribution shift* arises when an algorithm is trained on samples drawn from one probability distribution, but the distribution on which the algorithm is deployed—or tested—is different (Koh et al., 2021; Perdomo et al., 2020). As a result, the model that the algorithm learns based on the training distribution is unfit at test time. For instance, one would not expect an algorithm that is trained on criminal justice data in the U.S. to perform well for criminal justice decisions in the U.K. From the perspective of the initial training set, criminal cases in the U.K. look like exceptions.

4. *Partial observability* captures scenarios where not all of the relevant information is observable (Kalman, 1963; Kaelbling et al., 1998; Hashimoto et al., 2018). Suppose that two types of samples T_1 and T_2 exist, but the algorithm's input variables (a.k.a., features, covariates, attributes) are not rich enough to tell these samples apart. Suppose further that T_2 occurs less frequently than T_1. Then, an average-based model typically performs well on T_1 but not T_2 because, unable to tell them apart, the model lumps T_2 in with T_1 and treats them similarly. For instance, if a computer science department bases graduate admissions purely on an applicant's undergraduate major and GPA, it may not admit qualified applications who did not major in computer science but have relevant work experience after college.

Sampling bias, model (in)capacity, distribution shift, and partial observability show that there are many ways that exceptions appear in ML, so much so that the field has developed a language with which to discuss them.

Although the belief that averages are insufficient for many decision tasks is not new, there are few alternatives, and averages remain prominent in ML. Instead of ridding ourselves of averages, we must ask: *When are averages appropriate for our decision-making objectives and how should they be used?* Both decision makers and decision subjects are often satisfied with averages when a decision context is repeatable or has low stakes (Dawid, 2017; Parmigiani and Inoue, 2009). However, applying ML to repeatable or low-stakes decisions runs counter to the idea that the remarkable complexity and computational power afforded by ML would allow it to assist in decisions that are particularly challenging or one-off. As we continue to deploy ML as decision aids, this Article seeks to bring attention to the cases on which averages fail: the *exceptions.*

2.2 Dignity

Bringing attention to exceptions that are otherwise ignored by systems that work well for the majority has legal grounding. One influential concept that shifts attention from the aggregate to the individual is *dignity*.

Dignity is a concept that appears in international human rights law and domestic constitutions (O'Mahony, 2012). Despite being widely acknowledged as a "foundational principle", its meaning and consequent role in law remain unclear (O'Mahony, 2012; Rao, 2011; Glensy, 2011). At its core, dignity emphasizes the value of individuals. It has been used—in different and, at times, conflicting ways—to justify the right to free speech (Cohen v. California, 1971b), gay couples' right to marry (In re Marriage Cases, 2008), a woman's freedom to choose an abortion (Planned Parenthood of Southeastern Pa. v. Casey, 1992), and more. Its flexible meaning allows it to serve as a unifying theoretical basis for human rights and is part of the reason it appears in the Universal Declaration of Human Rights, which states that "all human beings are born free and equal in dignity and rights" (Assembly and Puybaret, 1999). Although there are multiple notions of dignity, we focus on two.

The first is the notion of *inherent dignity*, as popularized by Kant (2017), who states that all humans possess "a dignity (an absolute inner worth) by which he exacts respect for himself from all other rational beings in the world" and that this dignity cannot be substituted, exchanged, gained, or lost. Inherent dignity is based on the belief that, by virtue of being human, individuals must be afforded a "necessary respect" by others and the state (Gewirth, 1992). Kant (1967) also believed that individual autonomy and self-determination are special to humans and therefore intrinsically tied to dignity. In practice, inherent dignity is associated with negative liberty—a freedom from interference by the state that is rooted in the idea that a "person's dignity is best respected or enabled when he can pursue his own ends in his own way" (Rao, 2011).

The second notion of dignity relevant to this work is *dignity as recognition*, which requires that there be "esteem and respect for the particularity of each individual" (Rao, 2011). It demands that an individual's uniqueness is recognized and respected. Recall that inherent dignity is rooted in the idea that all individuals possess an inner worth that is deserving of respect regardless of whether their dignity is recognized. By contrast, under the concept of recognition dignity, an individual "can have dignity and a sense of self only through recognition by the broader society" (Rao, 2011). That all individuals inherently possess dignity is a "presumption of human equality" (Rao, 2011). On the other hand, dignity as recognition requires "treatment that *expresses* the equal worth of all individuals and their life choices" despite their differences (Rao, 2011). Rather than freedom from interference, recognition dignity is a positive concept in that the state must protect recognition dignity by enforcing respect between citizens and designing policies that actively acknowledge the equal worth of each individual (or group) in their uniqueness (Rao, 2011). In the past, recognition dignity has been invoked in claims against defamation and hate speech as well as the right to develop one's personality more broadly (Post, 1986; of Canada, 1990; Gazette, 2020).

The respect for an individual's uniqueness that is demanded by recognition dignity is at the core of this Article's right to be an exception. In highlighting how the reliance of data-driven decisions on generalizations can inflict harm on exceptions by no fault of their own, the right to be an exception "formalizes a basic respect for individual human dignity in a political system that otherwise allocates costs and benefits on the basis of majority rule" (Paradis, 2015). In this way, it can be viewed as a mechanism for protecting the recognition dignity of individuals in high-stakes, data-driven decision contexts. Recognizing the dignity of decision subjects does not require that decisions always tip in their favor. It simply requires a respect for dignity—an acknowledgment that when a decision can inflict significant harm on the subject, the decision should be based on a "respectful deliberation" that balances the subject's unique circumstances alongside other considerations (Harel, 2014). We will also see in Section 3 that the right to be an exception touches on other aspects dignity, including autonomy and self-determination.

One may then wonder why the right to be an exception is needed given the importance that both international and domestic law already place on dignity. Despite its similarity to the right to be an exception, a right to dignity in data-driven decisions is too abstract to be operational on its own. As thoroughly examined by Glensy (2011), Rao (2011), and O'Mahony (2012), the concept of dignity is so malleable that it can be invoked in many, often conflicting ways. Moreover, claims based on dignity "are most likely to succeed when coupled with an underlying deprivation of individual rights", especially in the U.S. where the constitutional structure emphasizes negative liberties, because the amount of respect one's dignity demands is highly subjective and context-dependent (Rao, 2011). As such, although dignity lays the foundations for the right to be an exception, the latter re-examines and refines it for data-driven decisions contexts.

2.3 Right to individualized sentencing

In the previous section, we observed that attention to exceptions is legally supported by the notion of dignity; in particular, recognition dignity, which demands that society recognize and respect the unique attributes that distinguish individuals. In this section, we turn to an example of a right that invokes dignity. Using the *right to individualized sentencing determinations* as a case study, we examine how exceptions arise and have been addressed by law.

In the 1970s, mandatory minimum sentences became more commonplace in the U.S. as part of an effort to "make sentencing procedures fairer and sentencing outcomes more predictable and consistent" (NRC et al., 2014). This change had unintended consequences, including shifting the power in sentencing determinations from the judge or jury to the prosecution who could, for instance, leverage mandatory minimum sentences to overcharge and subsequently obtain easy pleas (Berry III, 2019). During this time, many states adopted mandatory death penalties for felony convictions, but rather than improve the fairness and consistency of sentencing, mandatory death penalties reduced the ability of sentences to reflect the degree of *mens rea* revealed during trial (Berry III, 2019; Woodson v. North Carolina, 1976a; McGautha v. California, 1971a) It also forced the hand of juries, many of which refused to "convict murderers rather than subject them to automatic death sentences" (Woodson v. North Carolina, 1976a).

In 1978, the U.S. Supreme Court ruled in *Lockett v. Ohio* (1978) that defendants in capital cases are entitled to "individualized sentencing determinations" due to the seriousness and irrevocability of the death penalty (Furman v. Georgia, 1972). Specifically, the Court ruled that the Eighth Amendment prescribes a "fundamental respect for humanity" that "requires consideration of the character and record of the individual offender and the circumstances of the particular offense" before imposing a sentence as serious as the death penalty (Woodson v. North Carolina, 1976b). In 2012, the Court extended this concept to juvenile life-without-parole sentences in *Miller v. Alabama* (2012a), arguing that life-without-parole constitutes an especially serious sentence and that juvenile offenders are "constitutionally different from adults for purposes of sentencing" because "juveniles have diminished culpability and greater prospects for reform" (Miller v. Alabama, 2012b). More recently, Berry III (2019) has argued that individualized sentencing determinations should be broadened to all felony cases because felony convictions carry serious consequences. Berry III (2019) explains that felony convictions result in "dehumanizing effects that extend far beyond release, including the loss of right to vote, government surveillance, loss of possession and use of a firearm, housing consequences, employment consequences, and public benefits, not to mention social stigma" and therefore that the consequences of a felony conviction can be viewed as the "death of one's non-felony self".

The push for individualized sentencing determinations reflects a belief that, when the risk of harm is particularly high, a decision subject's unique circumstances and how the decision may affect them deserves careful consideration. In the words of Berry III (2019), sentencing determinations that are particularly serious "ought to include consideration of all relevant aggravating and mitigating evidence, and not flow automatically from the type of crime committed". Moreover, in practice, mandatory sentencing statutes re-delegate sentencing discretion from the court to the prosecution, and "prosecutorial decision making, by contrast, occurs in a black box of secrecy," which is exacerbated by the fact that "prosecutors enjoy a complete lack of accountability" in how they set the sentence for the offender (Berry III, 2019). Two parallels can be drawn from Berry's analysis. First, individualized sentencing echos the idea that there are conditions under which it is not appropriate to apply broad-strokes generalizations because using an aggregate rule fails to capture all cases and that these cases *matter*. By definition, the cases that fall through the cracks are exceptions. Individualized sentencing is therefore a movement towards greater consideration of exceptions when decisions have serious consequences. Second, rules based on generalizations can be misused. Applying a framework designed around general statements to the exception is like trying to fit a square into a circle, and the framework can consequently be exploited to the detriment of the individuals who do not fit neatly into the framework. Without sufficient transparency and accountability in sentencing determinations, broad-strokes generalizations not only fail on the exceptions, but they also fall short in their original goals of producing more fair procedures and more consistent, predictable outcomes.

3 The right to be an exception

In the early days of ML, algorithm designers could afford to write off exceptions because algorithms were evaluated on toy settings with low stakes. As ML becomes more ubiquitous and the stakes continue to rise, we must reconsider how exceptions are treated by data-driven decisions.

Exceptions can fall to the wayside for many reasons. Perhaps the data on which an algorithm is trained is not diverse or balanced enough to draw attention to the anomalous cases. Perhaps the model class is not rich enough to

be represent more than a few average-case rules. Perhaps the objective function explicitly optimizes for a rule that performs best on average, as it does in maximum likelihood estimation or empirical risk minimization. Whatever the reason, exceptions are often overlooked in ML, and in some cases, this oversight is dangerous.

In this Article, we argue that when a data-driven decision directly affects an individual whom we call the decision subject, the individual has the *right to be an exception*. By definition, exceptions are rare, and it is therefore impossible for every individual to be an exception. The right does not imply that every individual *is* an exception but that, when a decision may inflict harm on the decision subject, the decision maker should consider the possibility that the decision subject *may* be exception to the rule, choosing a decision that inflicts harm only when the decision maker is sufficiently confident in this decision. The greater the risk of harm, the greater the consideration. In this way, the right to be an exception does not prohibit the use of data-driven assessments in decision making. It serves as a safeguard for decision subjects by encouraging the final decision maker—algorithmic or human—to consider a decision's potential *harm*, the *suitability* of the data-driven assessment for the case at hand, and the *uncertainty* affecting the decision. This right would, for instance, require uncertainty be meaningfully incorporated by the algorithm or meaningfully communicated to a human decision-maker when the risk of harm is high.

In this section, we study what the right to be an exception means for both the decision subject and decision maker. We begin by walking through an example. We then discuss how the right to be an exception relies on three concepts: *harm*, *individualization*, and *uncertainty*. The risk of harm determines when it is necessary to afford an individual the right to be an exception. Individualization is important as it moves data-driven decisions away from generalizations and towards more fine-grained analyses tailored for each decision subject. Uncertainty is the final ingredient. When a decision may inflict harm on the decision subject, the uncertainty of each piece of evidence and each data-driven assessment determines whether the decision maker is confident enough to make a decision that inflicts harm. Importantly, the right to be an exception is not simply about obtaining more information. If so, addressing exceptions would be a matter of greater individualization. Rather, it also places emphasis on the roles uncertainty and harm play in decision-making. We conclude this section by discussing why the right to be an exception fills a gap in the governance of data-driven decision. We consider how the right adapts concepts intuitive in human decision-making to data-driven decision contexts, how the right gives individuals a path for legal recourse, and how it more fairly redistributes the burden of proof. We also touch on how establishing the right to be an exception can improve the legitimacy of data-driven decisions and mitigate issues associated with the scale and opacity of algorithms.

3.1 The right to be an exception through an example

In this section, we introduce the right to be an exception through an example; specifically, by studying the use of data-driven algorithms in criminal justice systems.

In the U.S., data-driven assessments have been used by judges as decision aids. For example, one such assessment produces a risk score that estimates a defendant's likelihood of recidivating—or re-offending—if the defendant is granted parole (Kehl and Kessler, 2017; Equivant, 2022). In recent years, judges have begun using risk scores for not only parole decisions, but also bail and sentencing. The appeal of these algorithms is their potential to reduce inconsistencies and bias in criminal justice decisions. Indeed, one study shows that the leniency of human judges depends on when cases are schedule, implying that the same defendant may receive different sentences should they be the first on a judge's docket rather than the last, an inconsistency that can be mitigated with algorithmic decision aids (Plonsky et al., 2021). Although data-driven algorithms have been shown to exhibit racial bias in criminal justice, a recent study demonstrates that, compared to humans, algorithmic decision aids would improve accuracy and reduce racial bias in pretrial decisions (Kleinberg et al., 2017).

Despite their promise, data-driven algorithms must be used with caution. Algorithms that rely heavily on averages may do so at the risk of washing out details that make each defendant's case unique and ignoring the defendant's capacity to be judged according to their own actions rather than the actions of others. Suppose, for instance, that a defendant is statistically similar to previous defendants who re-offended after released on parole. This statistical similarity on its own does not justify denying the defendant parole. For one, the defendant's unique circumstances may not be fully captured by the attributes used to compare defendants, making the statistical similarity compelling but imprecise, especially in the context of a decision that may cause the decision subject significant harm. For another, denying a defendant parole on the basis that similar defendants recidivated is a failure to respect, as Walen (2011) puts it, an individual's "autonomous moral agency". As articulated by Jorgensen (2021), it does not respect the "separateness of persons": it "treat[s] the wrongdoing by some as justification for imposing extra costs on others".

In other words, rather than be held responsible to their own actions, a defendant is made to pay for the actions of previous defendants.

Removing data-driven assessments from decision-making would not change the fact that historical data is used to inform decisions. After all, human judgments are also made on the basis of historical data; namely, the facts of the case and the judge's experiences. Then, where can a data-driven decision go wrong? There are two components to the answer: *harm* and *uncertainty*. Because the decision to deny parole inflicts significant harm on the decision subject, such a decision requires that the judge is given enough reason to believe the defendant will recidivate with high certainty. Therefore, if the uncertainty of a data-driven assessment is high and the risk of harm is great, then the assessment must be used with appropriate caution. However, serious consideration of harm and uncertainty is often omitted in data-driven decision contexts, as examined next.

3.1.1 Individualization is good but not enough

One way to improve data-driven assessments is greater individualization. We define individualization as the process of tailoring a data-driven assessment to the specific circumstances of a case. The greater the individualization, the more tailored the assessment. Individualization is an *information* concept in that one can only tailor an assessment to an individual if given enough information about that individual.

Much of ML is founded on the belief that, given enough information, a sufficiently individualized assessment can produce an estimate that matches the ground truth outcome. In the context of parole decisions, the belief is that, with enough information about the defendant and enough historical data, a risk assessment can predict whether the defendant recidivates with perfect accuracy. As a result, much attention has been paid to individualizing data-driven assessments, and these assessments are often justified based on their level of individualization.

However, while greater individualization is one piece of the puzzle—as it encourages fine-grained assessments—the right to be an exception would fall flat if it simply required a decision maker to claim that its data-driven assessment is sufficiently individualized. The missing piece is an acknowledgment of *uncertainty*. While individualizing the defendant's risk score is a necessary step to producing more reliable and accurate assessments, there are always sources of uncertainty, especially in high-stakes, one-off decision contexts. This uncertainty must be accounted for either by the data-driven algorithm or the decision maker who employs it, no matter how individualized the assessment is.

Recalling Walen's analysis, one source of uncertainty in parole decisions that cannot be removed with greater individualization is each defendant's moral autonomy. Simply put, denying parole on the basis of a highly individualized assessment only ensures that, instead of paying for the actions of a more general population of previous defendants, the current defendant pays for the actions of a subset of them, albeit a subset who bear close resemblance to the current defendant. No matter how individualized, a data-driven risk score *on its own* is a judgment of the defendant based on the actions of others rather the actions of the defendant. From a technical perspective, overfitting to previous defendants fails to uphold permutation invariance: the principle that, in general, the true likelihood that B will recidivate should not change whether A commits their crime before B (and A is therefore in the training set for B) or vice versa. From a philosophical perspective, arguing that, with enough individualization and historical data, a risk score can predict an outcome perfectly is a form of predeterminism: that whether the defendant recidivates is not their own choice but a function of the actions of previous defendants. All this to say that, because denying parole can inflict harm, a data-driven decision cannot rely on a risk score on its own, no matter how individualized. While an individualized assessment is better than one that is not, every assessment must be considered alongside the uncertainty it contains.

The right to be an exception captures precisely this notion. It can be viewed as the right to be given the opportunity to defy the rule. The right to be an exception does not imply that the decision subject *is* an exception but that they deserve to be considered as a possible exception when a decision could inflict significant harm on the decision subject. The greater the risk of harm, the more serious the consideration of uncertainty must be.

3.1.2 Balancing harm and uncertainty

Operationally, the right to be an exception in the context of parole decisions means that the decision maker's baseline belief, or null hypothesis, is that the decision subject is an exception if rejecting the null hypothesis would inflict harm. Walen (2011) expresses this sentiment as: "a state must normally accord its autonomous and accountable citizens [the] presumption [that they are law-abiding] as a matter of basic respect for their autonomous moral agency". Stated differently, a judge should presume that the defendant is law-abiding unless given sufficient evidence

to the contrary. One may then ask how much evidence (or certainty) is necessary to override a presumption that the defendant is law-abiding. Jorgensen (2021) offers the following:

> "It is morally negligent or reckless to intentionally harm someone unless we have not only reasonably high credence ... that the action is morally appropriate ... Very roughly: our present evidence must be such that little if any new information ... would cause our credence to drop below the threshold. The more harmful the interference, the more resilient the credence must be to justify it."

In the language of exceptions, Walen's and Jorgensen's arguments can be stated as follows. When a decision can inflict harm on the decision subject, the decision maker should give the subject an opportunity to be an exception. Even if an assessment suggests that the defendant will recidivate, unless it is with high certainty, the decision to reject the presumption that the defendant law-abiding should follow only if the judge's belief is so strong that very little if any new information would cause it to wane. One consequence of the right to be an exception in this context is that the uncertainty of a data-driven assessment must be meaningfully communicated to the judge.

3.2 Individualization, harm, and uncertainty

In the previous section, we examined how exceptions arise in parole decisions and explored how the right to be an exception can ensure that decision subjects receive appropriate consideration in the face of uncertainty and harm. In this section, we discuss the interplay between harm, individualization, and uncertainty in greater detail.

Recall that an exception is an instance that is excluded from a general statement or does not follow an expected rule. One may then be tempted to conclude that appropriately identifying and treating exceptions is simply a matter of *individualization*, which we refer to as the tailoring of a data-driven assessment to the specific circumstances of a case. For example, in the case of parole decisions, a data-driven assessment produces a risk score, and the more individualized the assessment, the more tailored the risk score is to the defendant's circumstances. Individualization addresses the problem of statistical stereotyping: the use of generalizations to reason about individuals. As discussed in Section 3.1.1, the belief behind individualization is that, with a sufficiently rich inputs, models, and data, one could obtain a perfect model whose assessment matches the ground truth outcome.

Individualization is indeed part of the answer to the question of how to better handle exceptions, as it moves data-driven assessments away from generalizations and towards more fine-grained analyses, but it is not the whole answer. As observed in the context of parole decisions, there are always sources of uncertainty—aside from random noise—that cannot be removed. While individualization gets the decision maker closer to estimating the outcome of interest for each individual decision subject, the decision maker must also confront whether the certainty behind this estimate is high enough to, together with all other relevant information, justify a decision that inflicts harm.

For example, an un-removable source of uncertainty in college admissions is how a student will perform when placed in a new environment, namely, the college under consideration. Even if the student is similar to previous students for which there is data on their performance at that college, not only could one argue that a student's performance is not predetermined (i.e., they have the ability to perform differently from past individuals), but one could also argue that there is a selection bias in that the previous students for which there is data are not a random sampling of students but students who were admitted to that college. As another example, consider a decision maker who must allocate medical resources across patients based on the patients' medical needs. The decision maker must therefore infer which patients have higher medical needs than others based on their condition, test results, medical history, age, race, and more. As demonstrated by Obermeyer et al. (2019) who showed that a data-driven algorithm exhibited significant racial bias when allocating medical care, there is always unknown information that may be relevant, whether due to noisy measurements, limited testing, overlooked confounders, or even that a medical condition has yet to be discovered.

The only way that an assessment is guaranteed to be perfect and rid of uncertainty is for the target variable itself to be input to the assessment, but this logic is circular. If one could measure the target variable, one would not need to infer it. In fact, Wolfram and Media (2002) has argued that some prediction tasks are computationally irreducible. Computational irreducibility is the theory that, from a computational point of view, some outcomes are simply unknowable (i.e., cannot be estimated or predicted perfectly). Because the process that produces the outcome can be as complex as the world itself, one could not possibly hope to perfectly estimate the outcome using an algorithm whose complexity is strictly lower than the complexity of the world that generated the outcome. Even if one had access to a model whose complexity is large enough to yield perfect performance on training and evaluation data,

one simply cannot be sure that it is rich enough and individualized enough to predict an outcome for an instance that it has never seen. Put simply, when it comes to sources of uncertainty, we cannot know what we do not know.

In most contexts, these sources of uncertainty are minor enough that individualization is sufficient in the limit. That is, with enough information, one can estimate the target with a level of accuracy that is sufficient for the decision context. However, when the decision is high-stakes, each source of uncertainty gains importance, and individualization may not be enough. This interplay between harm, individualization, and uncertainty is precisely what the right to be an exception captures. It demands that, when a decision can inflict harm on the decision subject, the decision maker should only choose to inflict harm when their level of certainty is high enough. The greater the risk of harm, the greater the required certainty. Because the right requires that the decision maker's credence be high when there is risk of significant harm, a decision maker cannot ignore the uncertainty that accompanies a data-driven assessment. Therefore, the right to be an exception affirms that individualization is critical but only part of the answer. It helps us get closer to an ideal model whose assessments resemble the ground truth, but there are always sources of uncertainty, and these sources *matter* when the decision is high-stakes.

And perhaps obtaining a perfect model is besides the point. As we saw above, every model contains some degree of uncertainty. The right to be an exception shifts attention away from finding a perfect model and forces the decision maker to consider harm and uncertainty alongside individualization. It can therefore be understood as ensuring that, when following a rule may inflict harm on the decision subject, the decision maker must think seriously about their level of certainty by considering the possibility that the decision subject just *may* be an exception to the rule.

3.3 The gap the right to be an exception fills

The right to be an exception fills an important gap in the governance of data-driven decisions. As discussed in Section 2.2, although dignity underlies the right to be an exception, dignity is too malleable a concept for our purposes. Furthermore, although there exist the right to individualized sentencing and similar notions, the unique challenges posed by ML create opportunities for significant and widespread harm that our legal systems are ill-equipped to handle. As stated by Kaminski and Urban (2021), "[artificial intelligence] warrants unique risks that deserve distinct treatment". In this section, we unpack the ways the right to be an exception fills a gap in the governance of data-driven decisions.

3.3.1 Legal counterweight to tendency of data-driven decisions to rely on averages

The right to be an exception provides decisions subjects a path for legal recourse when data-driven decisions do not appropriately consider individualization, harm, and uncertainty. Although considering these three factors is natural to human decision makers, decisions that are informed or made by algorithms often overlook this balancing act.

One reason why exceptions may be overlooked is that averages are so foundational to ML. Averages appear in every part of the data-driven decision-making pipeline. They can appear in the stated objective. For example, maximum likelihood estimation and empirical risk minimization are both notions of averages. They can also appear in evaluation through metrics like accuracy, which is a measure of the average performance across the population represented by the evaluation dataset. Although high accuracy is often taken as an indication of the confidence one should place in a model, high accuracy implies good performance on average and not necessarily that the model will perform well on a specific decision. Even proposals for individual-level rights in data-driven decision contexts mistake metrics like accuracy for an indication of a model's suitability (Wachter and Mittelstadt, 2019), as was the case in *State v. Loomis* (2016) discussed below.

The message is not that we must abandon averages or find the "perfect" metric but that data-driven assessments must be viewed for what they are. In particular, when an algorithm is designed to work well in aggregate, its assessments should be taken with a grain of salt when the decision-maker cares about the outcomes of *individuals*. As such, the right to be an exception helps both the decision subject *and* the decision maker. It gives the decison subject assurance that the decision is sufficiently individualized to their circumstances with an appropriate consideration of harm and uncertainty. It also gives the decision maker the ability to move beyond focusing on whether they have chosen the "best" model with the "best" objective function and "best" performance according to the "best" metric. In this way, the right to be an exception extends a concept intuitive to human decision-makers and legally substantiated by recognition dignity, expressing it explicitly for data-driven decision contexts.

3.3.2 Legal recourse and legitimacy

Establishing the right to be an exception is the first step in providing individuals with a path to *contest* data-driven decisions that rely heavily on broad-strokes generalizations. The ability to contest ML systems is highly important. Instead of trusting on faith alone that data-driven decisions are suitable for every context on which they are applied, the ability to contest gives individuals *agency* over the decisions that affect their lives (Kaminski and Urban, 2021). In doing so, "a fair contestation process can enhance the perceived *legitimacy*" of and *trust* in data-driven decisions, to the benefit of both decision subjects and makers (Kaminski and Urban, 2021). On one side, algorithm designers can continue to develop and deploy. On the other side, decision subjects can serve as checks, ensuring that the algorithms are applied appropriately by helping to identify when they fail. In this way, contestation serves as a "systematic management technique" that smooths the integration ML into decision-making by "uncovering errors, identifying their causes, and providing schemes and incentives to correct them" (Crawford and Schultz, 2014).

Affording individuals the ability to contest data-driven decisions on the basis of the right to be an exception does not mean that data-driven assessments should be abandoned. The right serves as a legal complement to efforts by the computer science community to improve performance on all cases, including the exceptions. Rather than bar data-driven assessments in decision-making, contestation identifies directions for algorithm improvement, "provid[ing] individuals recourse even when they choose to continue to participate in the activity" (Kaminski and Urban, 2021).

A fair contestation process also articulates the level of *transparency* an algorithm should exhibit. In the absence of perfect transparency—a concept that continues to elude us because algorithms, even with full access, can be difficult to intuit—contestation serves as a rigorous alternative that sets guidelines for how transparent a system must be, which is to say as transparent as needed to determine whether it upholds an individual's rights.

3.3.3 Burden of proof

By shifting power away from decision makers who are currently permitted to argue that excellent (or even good) average-case performance justifies the poor treatment of the few exceptions, the right to be an exception also re-balances the burden of proof. At the moment, plaintiffs "face an uphill battle ... with regards to big data inferences" because the legal system is not designed to handle the unique challenges of ML (Kaminski and Urban, 2021; Ajunwa, 2021; Barocas and Selbst, 2016). Unlike with humans, one cannot base a legal claim on the algorithm's "intent" (e.g., racial animus), and it is often difficult to gain insights into an algorithm's assessment. In place of these normal routes, algorithms are often justified using evidence of good performance, such as high accuracy. However, as noted in this Article, most common metrics, including accuracy, are *averages* that do not indicate that an assessment performs well on an individual case, making them unsuitable in high-stakes, non-repeatable contexts. Therefore, when individuals contest data-driven decisions, they must argue that an algorithm is unsuitable on the basis of averages, which tends to fall in favor of the decision-maker. The right to be an exception redistributes this burden, requiring that decisions at risk of inflicting significant harm demonstrate adequate attention to individualization, harm, and uncertainty.

As an example, consider *State v. Loomis* (2016). In 2013, Eric Loomis was charged in relation to a drive-by shooting. Although Loomis denied participating in the shooting, he conceded to driving the same car that day and pleaded guilty to "attempting to flee a traffic officer and operating a motor vehicle without the owner's consent," two of the less severe charges (State v. Wisconsin, 2016). An algorithmic risk assessment, whose methodology is a trade secret, was consulted as a part of his sentencing determination, and Loomis was sentenced to six years of imprisonment and five years of extended supervision. Loomis filed for post-conviction relief. He argued (a) that the use of an algorithmic risk assessment infringed on both his right to an individualized sentence and his right to be sentenced on accurate information; and (b) that the risk assessment—and thereby the court—unconstitutionally used gender to determine his sentence.

Loomis was denied post-conviction relief on two grounds. First, that the use of gender "served the nondiscriminatory purpose" of improving the accuracy of the algorithm (State v. Wisconsin, 2016). Second, because the risk assessment uses only publicly available data and information about the defendant, Loomis "could have denied or explained any information that went into making the report and therefore could have verified the accuracy of the information used in sentencing" (State v. Wisconsin, 2016). We contend that the court made several missteps in their ruling. For one, by claiming that gender serves a "nondiscriminatory purpose", the court appealed to the concept of intent, a concept lacks meaning when assigned to algorithms. For another, using accuracy to justify the algorithm prevented Loomis from receiving more serious consideration about whether an *aggregate* measure like accuracy was suitable for his *specific* case. Finally, arguing that Loomis could have verified the algorithm's accuracy because all

the data is available to him placed an enormous burden of proof on the plaintiff, who is generally not expected to know how a risk assessment produces predictions from data.

Although the court recommended that judges should be provided with several warnings explaining the shortcomings of risk scores, some have argued that, unless algorithm designers are expected to provide more meaningful information about the assessments, "judges will not be able to calibrate their interpretations" and cannot know "how much to discount these assessments" (Harvard Law Review, 2017). To this end, the right to be an exception would require that the suitability of a data-driven assessment that may inflict harm and the uncertainty it carries are meaningfully incorporated into the decision or meaningfully communicated to the decision-maker. In this way, the right to be an exception shifts the large burden of proof currently placed on decision subjects toward decision makers, who would be compelled to demonstrate that a data-driven assessment allows appropriate consideration of individualization, harm, and uncertainty.

3.3.4 Scale and opacity

We conclude with two final notes. The first is that, due to the scale at which ML is being applied, the right to be an exception may serve as a legal brake to data-driven *feedback loops*. These feedback loops occur because ML can be applied at large scales and with great efficiency, often to the detriment of a subset of the population. For example, suppose that a welfare algorithm performs well on the large majority, but poorly on a small number of individuals whom it mistakenly classifies as high-risk for welfare fraud. If this algorithm is used by many welfare agencies, then the same individuals are consistently denied welfare. Because the individuals that an agency chooses to grant welfare are eventually used to train future algorithms, these individuals will continue to be classified under "reject", leading to a feedback loop. The right to be an exception can slow—and hopefully stop—such feedback loops by requiring that decision-makers pay greater attention to the exceptions. Notably, the right to be an exception is not subsumed by rights that protect against discrimination on the basis of sensitive characteristics because, due to the complex and unintuitive nature of algorithms, the exceptions missed by algorithms do not fall neatly along demographic lines. As our second and final note, the uncertainty component of the right to be an exception may mitigate challenges posed by the inability to access or intuit algorithmic logic, also referred to as *algorithmic opacity*. When humans provide assessments, a decision maker is often able to detect intent (e.g., racial animus) and places a degree of trust in the assessment accordingly. Although algorithms lack intent, one of the primary purposes of detecting intent is to assign credibility to an assessment. In the absence of intent, the right to be an exception encourages the decision maker to incorporate assessments with appropriate caution and skepticism, effectively replacing a credibility assignment.

4 Operationalizing the right to be an exception

In this section, we examine how one could operationalize the right to be an exception. We consider ex ante and ex post legal measures as well as potential technical approaches.

4.1 Legal measures

4.1.1 Ex ante

The ex ante measures of the right to be an exception would require that a data-driven decision appropriately considers the three main components of the right: harm, individualization, and uncertainty.

Specifically, the data-driven assessment must (1) evaluate the potential harm that the decision could inflict; (2) justify the assessment on the basis of its level of individualization; and (3) demonstrate that, give the level of harm and individualization, the assessment appropriately and meaningfully incorporates uncertainty or appropriately and meaningfully communicates it to the human decision maker.

To determine harm for (1), one can use the standard of "significant effects" established in Article 22(1) of the General Data Protection Regulation (2016), whose scope has also been the subject of study (Kaminski and Urban, 2021). For (2), it currently suffices to justify a data-driven assessment on the basis of aggregate measures, such as accuracy. Even those calling for individual-level rights in data-driven decision contexts recommend that only assessments that are "accurate and statistically reliable"—both of which are aggregate notions—are used, illustrating the need for a legal standard or rule that ensures data-driven assessments are sufficiently individualized. One possibility would be to require that data-driven algorithms report performance on more fine-grained metrics—such as multicalibration

(Hébert-Johnson et al., 2018)—in addition to aggregate ones. Finally, meaningfully incorporating or communicating uncertainty for (3) is an active area of research in human-computer interaction (Hullman, 2016; Hofman et al., 2020). One way to do so would be to *a priori* identify the sources of uncertainty that would be informative for the given decision context and require that the assessment provide uncertainty estimations for each, similarly to how existing works distinguish and report on epistemic and aleatoric uncertainty (Kendall and Gal, 2017).

4.1.2 Ex post

The ex post component of the right to be an exception is contestation. As explained in Section 3.3.2 and explored by Kaminski and Urban (2021), contestation is an accountability mechanism that enhances the legitimacy of data-driven assessments as well as builds the public's trust in them. The procedure for contestation may depend on the country of jurisdiction. As a model for the right to be an exception, one could turn to the procedure for contesting on the basis of Title VII of the U.S. Civil Rights Act's principle of disparate impact (Barocas and Selbst, 2016). In disparate impact cases, a plaintiff must first establish that an employment practice causes disparate impact with respect to a protected class, which can be countered if the defendant shows that the employment practice is rooted in "business necessity", which the plaintiff can then refute by providing an alternate employment practice that would mitigate disparate impact without violating business necessity. Contestation for the right to be an exception could mirror this procedure as follows. First, the plaintiff must establish that (1), (2), and/or (3) from Section 4.1.1 were violated by the data-driven decision. If the plaintiff is successful, the defendant can counter by showing that the data-driven decision could not have been changed without demanding significant resources or inflicting disproportionate harm on other parties. Finally, if the defendant is successful, the plaintiff can refute the defendant's justification by providing an alternate procedure that improves upon the assessment with respect to (1)-(3) and does not demand excessive resources or inflict disproportionate harm on other parties. This procedure is one among many possibilities, as surveyed by Kaminski and Urban (2021).

4.2 Technical measures

The notion of exceptions is not new in computer science. We conclude this Article by reviewing three areas of work that seek to better identify and treat exceptions. In addition to the three mentioned below, there are many other fields, including reinforcement learning and extreme event prediction, that are motivated by the belief that exceptions matter.

4.2.1 Causal inference

In recent years, there has been renewed interest in moving beyond learning correlation to instead determining causation (Pearl and Mackenzie, 2018; Wachter et al., 2017). The interest in cause-and-effect can be traced to the same intuition discussed in Section 3.2: when a decision may inflict harm and the circumstances are unique, a decision-maker is interested in understanding the sequence of events that leads to an outcome in order to make an informed decision under the current state of affairs. If, for instance, the set of conditions for a training example Z_1 (e.g., the qualifications of one job candidate) is identical to the set of conditions for another example Z_2 (e.g., the qualifications of another job candidate), the decision-maker may still be inclined to make a different decision for Z_2 than for Z_1 because the sequence of events that led to Z_1 may be different from the sequence of events that led to Z_2 (e.g., the paths that the two job candidates underwent in acquiring identical qualifications may be different, which may indicate that one candidate is better suited for the job than the other).

At its core, causal inference seeks to shift focus away from frequency analyses. Instead of learning relationships that depend on how often an outcome has occurred in the past, one seeks to learn the factors that led to an outcome. For example preparing an autonomous vehicle to avoid accidents can be distilled down to the idea that, even though accidents are rare, they are important events, and the system designer wishes to detect conditions that may lead to an accident and avoid actions that may cause an accident. In this way, causal inference is a mechanistic approach to appropriately identifying and treating outcomes of interest, including exceptions.

4.2.2 Robust optimization

Robust optimization is a broad field of study that aims to find reward-maximizing or cost-minimizing solutions that are robust under uncertainty (Bertsimas et al., 2011, 2018). The justification for robust optimization mirrors the

motivation behind thinking beyond averages. Because robust optimization encourages decision-makers to be more conservative, it is typically only applied when a decision is high-stakes. A canonical example for robust optimization is the design of a bridge. If the architect only minimizes the expected (i.e., average) cost to build—without considering possible defects in building materials, rare weather conditions, or mistakes during the building process—then the bridge may fail in unexpected ways. Although such outcomes are rare, they would cost lives. On the other hand, if the decision is low risk, then the decision-maker may be willing to take a risk if it results in lower costs.

Although many applications of robust optimization use averages (e.g., expected risk) in the objective function, introducing robustness shifts a decision's focus away from averages by placing emphasis on the presence of uncertainty. For example, consider a judge deciding whether to grant or deny parole to a defendant. One approach would be to maximize accuracy: to ensure that, on average, those granted parole are those that do not recidivate, and vice versa. A robust optimization approach may state that, due to uncertainty in the information on which the decision is made, when the predicted likelihood of recidivism is sufficiently small and the associated crime is minor, then maximizing expected accuracy is not enough of a reason to deny parole. Robust optimization is typically enforced via a constraint or minimax optimization. Intuitively, this approach defines a range of possibilities that must be considered and minimizes the worst-case risk within this range or maximizes some measure of performance while requiring that the worst-case risk is sufficiently low.

4.2.3 Algorithmic fairness

One of the first approaches to algorithmic fairness studies fairness as a group notion. The intent of group fairness is to provide a mathematical analog to legal notions of fairness, such as those outlined in Title VII of the U.S. Civil Rights Act, which protects against discrimination in hiring. Group fairness tackles fairness by requiring (approximate) parity in group-level statistics (e.g., equalized true positive rates across race, gender, or age).

While group fairness has led to many interesting research insights (Barocas et al., 2018; Chouldechova, 2017), there are examples in which group fairness is insufficient (Dwork et al., 2012; Ilvento, 2019). For example, under group fairness, it is possible for a hiring algorithm to select the "top" half of white applicants and the "bottom" half of black applicants and qualify as fair with respect to race because, for both white and black applicants, 50 percent are hired. At its core, group fairness is a notion of averages, and the goal is to achieve parity in proportions. Individual fairness rose as a counterpoint to this average-based notion of fairness (Dwork et al., 2012). Stated informally, individual fairness is the idea that similar individuals should receive similar treatment. It asserts that attention to parity across demographics should not pull attention away from the unique attributes of each individual. The example above does not satisfy individual fairness because it flips the way qualified White and Black applicants are treated. In this way, individual fairness seeks to shift focus away from averages, and areas such as intersectional fairness are on a similar mission. Numerous others notions of fairness have been proposed, particularly in economics. Minimax fairness minimizes the worst-case (i.e., maximum) harm across the units of interest, envy-freeness ensures that no unit would (subjectively) prefer the outcome of any other unit over their own, multicalibration requires calibrated predictions on every sub-population of interest, and many more context-dependent definitions of fairness have emerged (Budish, 2011; Bouveret and Lemaître, 2016; Hébert-Johnson et al., 2018). The area of algorithmic fairness approaches exceptions by directly measuring how decisions differentially impact the units of interest. As the size of these units decreases (e.g., from racial groups to intersectional groups to individuals), fairness provides increasingly strong protections for exceptional cases.

References

General data protection regulation (gdpr), 2016.

Wisconsin supreme court requires warning before use of algorithmic risk assessments in sentencing. *Harvard Law Review*, pages 1530–1537, March 2017.

Ifeoma Ajunwa. The Auditing Imperative for Automated Hiring. *Harvard Journal of Law & Technology*, 34(2):621–699, 2021.

United Nations. General Assembly and E. Puybaret. *Universal Declaration of Human Rights*. Inter-American Institute for Human Rights, 1999.

Solon Barocas and Andrew D Selbst. Big Data's Disparate Impact. *California Law Review*, 104:671, 2016.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and machine learning: Limitations and opportunities, 2018.

William W Berry III. Individualized Sentencing. *Washington & Lee Law Review*, 76:13, 2019.

Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, 2018.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006. ISBN 9780387310732.

Sylvain Bouveret and Michel Lemaître. Characterizing conflicts in fair division of indivisible goods using a scale of criteria. *Autonomous Agents and Multi-Agent Systems*, 30(2):259–290, 2016.

Eric Budish. The combinatorial assignment problem: Approximate competitive equilibrium from equal incomes. *Journal of Political Economy*, 119(6):1061–1103, 2011.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

National Research Council, Division of Behavioral & Social Sciences & Education, Committee on Law & Justice, and Committee on Causes & Consequences of High Rates of Incarceration. *The Growth of Incarceration in the United States: Exploring Causes and Ponsequences*. National Academies Press, 2014. ISBN 9780309298018.

U.S. Supreme Court. Mcgautha v. california, 402 u.s. 183, 199. 1971a.

U.S. Supreme Court. Cohen v. California, 403 u.s. 15. 1971b.

U.S. Supreme Court. Furman v. georgia, 408 u.s. 238, 290. 1972.

U.S. Supreme Court. Woodson v. north carolina, 428 u.s. 280. 1976a.

U.S. Supreme Court. Woodson v. north carolina, 428 u.s. 280, 304. 1976b.

U.S. Supreme Court. Lockett v. ohio, 438 u.s. 586. 1978.

U.S. Supreme Court. Planned parenthood of southeastern pa. v. casey, 505 u.s. 833. 1992.

U.S. Supreme Court. Miller v. alabama, 567 u.s. 460, 489. 2012a.

U.S. Supreme Court. Miller v. alabama, 567 u.s. 460, 471. 2012b.

Kate Crawford and Jason Schultz. Big data and due process: Toward a framework to redress predictive privacy harms. *Boston College Law Review*, 55:93, 2014.

Gary Cuddeback, Elizabeth Wilson, John G Orme, and Terri Combs-Orme. Detecting and statistically correcting sample selection bias. *Journal of Social Service Research*, 30(3):19–33, 2004.

Philip Dawid. On individual risk. *Synthese*, 194(9):3445–3474, 2017.

Luc Devroye, Lázló Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Stochastic Modelling and Applied Probability. Springer New York, 1997. ISBN 9780387946184.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Equivant. Software for justice, January 2022.

Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

Federal Law Gazette. *Grundgesetz Für die Bundesrepublik Deutschland [Basic Law for the Federal Republic of Germany]*. 2020.

Alan Gewirth. Chapter 1: Human dignity as the basis of rights. In Michael J. Meyer and William A. Parent, editors, *The Constitution of Rights*, pages 10–28. Cornell University Press, 1992.

Rex D. Glensy. The Right to Dignity. *Columbia Human Rights Law Review*, 43:1–65, 2011.

Alon Harel. *Why Law Matters*. Oxford Legal Philosophy. Oxford University Press, 2014. ISBN 9780191030734.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

Jake M Hofman, Daniel G Goldstein, and Jessica Hullman. How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

Xia Hu, Lingyang Chu, Jian Pei, Weiqing Liu, and Jiang Bian. Model Complexity of Deep Learning: A Survey. *arXiv preprint arXiv:2103.05127*, 2021.

Jessica Hullman. Why evaluating uncertainty visualization is error prone. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pages 143–151, 2016.

Christina Ilvento. Metric learning for individual fairness. *arXiv preprint arXiv:1906.00250*, 2019.

Renée Jorgensen. Algorithms and the Individual in Criminal Law. *Canadian Journal of Philosophy*, page 1–17, 2021. doi: 10.1017/can.2021.28.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Rudolf Emil Kalman. Mathematical description of linear dynamical systems. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 1(2):152–192, 1963.

Margot E Kaminski and Jennifer M Urban. The right to contest ai. *Columbia Law Review*, 121(7):1957–2048, 2021.

Immanuel Kant. *Grundlegung zur Metaphysik der Sitten*. Universal-Bibliothek. Reclam, 1967.

Immanuel Kant. *The Metaphysics of Morals*. Cambridge Texts in the History of Philosophy. Cambridge University Press, 2017. ISBN 9781107086395.

Danielle Leah Kehl and Samuel Ari Kessler. Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. 2017.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics*, 133(1):237–293, 08 2017. ISSN 0033-5533. doi: 10.1093/qje/qjx032.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. doi: 10.1126/science.aax2342.

Supreme Court of California. In re marriage cases, 43 cal.4th 757, 76 cal. rptr. 3d 683, 183 p.3d 384. 2008.

Supreme Court of Canada. R. v. keegstra, 3 s.c.r. 697. 1990.

Supreme Court of Wisconsin. State v. wisconsin, 881 n.w.2d 749, 754, 757. 2016.

Conor O'Mahony. There is no such thing as a right to dignity. *International Journal of Constitutional Law*, 10(2): 551–574, 03 2012. ISSN 1474-2640. doi: 10.1093/icon/mos010.

Michel Paradis. Dignity through Law. The New Rambler, December 2015.

Giovanni Parmigiani and Lurdes Inoue. *Decision Theory: Principles and Approaches*. Wiley Series in Probability and Statistics. Wiley, 2009. ISBN 9780470746677.

J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018. ISBN 9780465097616.

Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

Ori Plonsky, Daniel L Chen, Liat Netzer, Talya Steiner, and Yuval Feldman. Best to be last: Serial position effects in legal decisions in the field and in the lab. *Bar Ilan University Faculty of Law Research Paper*, (19-15), 2021.

Robert C Post. The social foundations of defamation law: Reputation and the constitution. *California Law Review*, 74:691, 1986.

Neomi Rao. Three Concepts of Dignity in Constitutional Law. *Notre Dame Law Review*, 86:1–183, 2011.

Wilko Schwarting, Javier Alonso-Mora, and Daniela Rus. Planning and Decision-making for Autonomous Vehicles. *Annual Review of Control, Robotics, and Autonomous Systems*, 1:187–210, 2018.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550 (7676):354–359, 2017.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning series. MIT Press, 2nd edition, 2018. ISBN 9780262039246.

Vladimir Vapnik, Esther Levin, and Yann Le Cun. Measuring the VC-dimension of a learning machine. *Neural computation*, 6(5):851–876, 1994.

Sandra Wachter and Brent Mittelstadt. A right to reasonable inferences: re-thinking data protection law in the age of big data and ai. *Columbia Business Law Review*, page 494, 2019.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Alec Walen. A punitive precondition for preventive detention: Lost status as a foundation for a lost immunity. *San Diego Law Review*, 48:1229, 2011.

S. Wolfram and Wolfram Media. *A New Kind of Science*. Wolfram Media, 2002. ISBN 9781579550080.