Outliers Exist: What Happens if You are a Data-Driven Exception?

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Data-driven tools are increasingly used to make consequential decisions. In recent years, they have begun to advise employers on which job applicants to interview, judges on which defendants to grant bail, lenders on which homeowners to give loans, and more. In such settings, different data-driven rules result in different decisions. The problem is, for every data-driven rule, there are exceptions. While a data-driven rule may be appropriate for some, it may not be appropriate for all. In this piece, we argue that existing frameworks do not fully encompass this view. As a result, individuals are often, through no fault of their own, made to bear the burden of being data-driven exceptions. We discuss how data-driven exceptions arise and provide a framework for understanding how we can relieve the burden on data-driven exceptions. Our framework requires balancing three considerations: individualization, uncertainty, and harm. Importantly, no single consideration trumps the rest. We emphasize the importance of uncertainty, advocating that decision-makers should utilize data-driven recommendations only if the levels of individualization and certainty are high enough to justify the potential harm resulting from those recommendations. We argue that data-driven decision-makers have a duty to consider the three components of our framework before making a decision, and connect these three components to existing methods.

1 Introduction

We make sense of our world through rules. But, to every rule, there are exceptions. Although exceptions are by definition uncommon, they often carry significance disproportionate to their numbers. Exceptions not only improve our understanding of the rules, but they also help us develop better ones. No matter how good the rule, it cannot work for everyone, begging the question: *What happens to the individuals for which a decision rule is unfit: the exceptions?*

In some cases, nothing. We accept that rules and generalizations are, on occasion, tolerable and even necessary (Lippert-Rasmussen, 2011). Indeed, the law allows landlords to put no-pet clauses in rental agreements (a rule based on the generalization that renters with pets cause more damage to homes than renters without pets) and airlines to remove passengers for safety reasons (a policy that relies on judgments about actions that a passenger could but has not yet committed).

In other cases—typically, when the risk of harm is high—the state steps in to shield individuals from the adverse effects that can follow from the over-application of rules. Consider sentencing decisions. For many crimes, there are mandatory minimum sentences: a set of standardized rules that prescribe the minimum sentence a defendant must serve for a crime, if convicted. These rules arose in the U.S. as a way to "make sentencing procedures fairer and sentencing outcomes more predictable and consistent" (Travis et al., 2014). Importantly, mandatory minimum sentences were also used in capital cases, i.e., cases in which the crime is punishable by death. In 1976, however, the U.S. Supreme

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

Court ruled in *Woodson v. North Carolina* that mandatory minimum sentences should *not* be applied to capital cases. The Court wrote that there must be "consideration of the character and record of the individual offender and the circumstances of the particular offense" before imposing a sentence as serious and irrevocable as the death penalty (U.S. Supreme Court, 1976). In other words, the Court decided that, when it comes to the death penalty, rules that regularly yield exceptions—in this setting, defendants on which the rule, but not the presiding judge, would impose the death penalty—are unacceptable. The courts responded by giving greater discretion to judges.

In this piece, we turn our attention to *data-driven rules*. By "data-driven rules," we refer to the decision rules behind data-driven decision aids. For example, data-driven decision aids in lending advise lenders on whether or not to grant a loan. Generally speaking, these aids produce a score for each applicant that predicts the likelihood that the applicant, if approved, would repay the loan. Different rules generate different scores. While one rule may give higher scores to applicants with families, another may not. One rule may use the applicant's zip code as an input while another may not. As such, an applicant may be approved for a loan under some rules but not others.

As many scholars have acknowledged, there is a gap in the governance of data-driven decisions because individuals who are subject to data-driven decisions are not always protected by a legal system that has been built around human decisions (Citron, 2007; Wachter and Mittelstadt, 2019; Kaminski and Urban, 2021). In this piece, we focus on what happens to individuals for which a data-driven rule is unfit. We find that, under existing frameworks, *individuals on which a data-driven rule fails are, through no fault of their own, made to bear the burden of those mistakes*. There are two characteristics of data-driven rules that exacerbate this problem. First, data-driven rules are highly non-intuitive. Second, they are frequently updated, oftentimes without any visible changes to those that employ them.

This setting stands in stark contrast to how we approach rules in our legal system. As opposed to standards, rules provide a clear delineation between behaviors that are legal and those that are illegal. Rules are furthermore public and relatively static. That is, whether or not individuals agree with rules that appear in the law, they are aware of the consequences of their actions, and they do not expect that these rules change without their knowledge. Data-driven decision-subjects, on the other hand, are often left in the dark about why they receive the decision that they did, and we show that they face an disproportionately high barrier to proving that the data-driven rule is unfit for them.

In this piece, we unpack this phenomenon in detail, and propose a framework that remedy the problem of data-driven exceptions.

2 Background

Bringing attention to exceptions that may be neglected by systems that work well for the majority has philosophical and legal grounding. One influential concept that emphasizes the importance of giving appropriate consideration and respect to each 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). It has been used—in different and, at times, conflicting ways—to justify the right to free speech (U.S. Supreme Court, 1971), a gay couple's right to marry (Supreme Court of California, 2008), a woman's freedom to choose an abortion (U.S. Supreme Court, 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" (United Nations General Assembly, 1948). Although there are multiple notions of dignity, we focus on two.

The first is the notion of *inherent dignity*, as popularized by Kant, 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 (Kant, 2017). 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 piece 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 (Post, 1986; Supreme Court of Canada, 1990; Federal Law Gazette, 2020).

The respect for an individual's uniqueness that is demanded by recognition dignity is closely related to our work. In highlighting how the reliance of data-driven decisions on rules can inflict harm on exceptions, our work emphasizes "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).

2.1 A Note on Related Works

There are many existing works on data-driven technologies and their pitfalls. These works have covered enormous ground, highlighting issues that arise during the application of data-driven technologies and gaps in their governance. We build on this literature, but there are four factors that together make this piece distinct.

First, many works (such as those examining disparate impact (Barocas and Selbst, 2016)) focus on group-based outcomes, e.g., discrimination based on a protected attribute. In contrast, our work examines data-driven decisions through the lens of individual outcomes rather than group-based ones. In particular, we discuss how one can determine if a data-driven rule is appropriate for a specific decision-subject, similarly to Lippert-Rasmussen (2011).

Second, most existing works propose to improve outcomes by requiring that data-driven tools be "fair," "accurate," and "reliable" (Citron and Pasquale, 2014; Wachter and Mittelstadt, 2019). We find that such criteria are important but do not capture the full picture when evaluating the suitability of a data-driven tool for a *specific* decision context. Accuracy, for instance, is an average notion—high accuracy only implies good performance in an average sense. Similarly, reliability implies good performance in a repeated sense—that, if run many times, an algorithm would consistently perform well. In this piece, we offer an additional consideration. In addition to fairness, accuracy, and reliability, there is another desideratum: a decision-maker should not presume that the data-driven rule is suitable for an arbitrary decision-subject, particularly when the stakes are high. Rather, the decision-maker should only apply a data-driven rule if they are sufficiently confident (as measured against the risk of harm) that it is indeed suitable, as detailed in Section 3.

Third, there are several works that examine whether data-driven rules should be sufficiently individualized in order to be applied. That is, they investigate how individualization addresses the problem of statistical discrimination (Lippert-Rasmussen, 2011; Wachter et al., 2021). In this piece, we build on this discussion and argue that individualization is one, but not the only, component of evaluating a data-driven rule's suitability. We maintain that one must also consider the data-driven rule's *uncertainty*, a concept that is often overlooked but is core to the our work.

Lastly, our hope is to provide a framework that can serve as a common language with which to discuss data-driven exceptions across disciplines. To this end, we also consider the technical aspects of data-driven exceptions, including their origins (showing that data-driven exceptions arise in more ways than existing works typically consider, making the problem less straightforward than commonly assumed) and the technical viability of the proposed solutions. We pay particular attention to the latter. For example, although open-sourcing data-driven tools may be useful, it is infeasible in many cases (e.g., due to trade secret law or that open-sourcing introduces vulnerabilities to adversarial attacks). To ensure technical viability, we distill the our framework down to three concepts, described in Section 3, that are also meaningful in machine learning.

3 Proposed Framework

In this piece, we argue that, a decision-maker cannot presume that a data-driven rule is suitable for a given decision-subject—they must be sufficiently confident (relative to the risk of harm) that the individual is not an exception. In other words, a data-driven decision-maker—whether a machine or machine-aided human—must make a decision that inflicts harm only if they have applied due care and due diligence in determining whether the data-driven rule is fit for the decision-subject in question. The greater the risk of harm, the higher the bar.

Society has, for the most part, developed standards for assessing whether a human has applied due care and due diligence in decision-making (cf. the right to individualized sentencing (Berry III, 2019)). After all, the law has been honed to work for human-driven decisions. How one would operationalize this concept in the data-driven context is, however, unclear. In this piece, we propose that adapting this requirement for data-driven decisions can be achieved by considering three factors: *individualization*, *harm*, and *uncertainty*. Via these three components, we provide a concrete framework through which a decision-maker can determine when a data-driven rule is appropriate or a decision-subject can determine whether to contest a data-driven decision. Importantly, no components on its own is sufficient (or necessarily even desirable). For instance, as we unpack below, individualization can often give rise to undesirable effects.

3.1 Individualization: Moving from the Aggregate to the Individual

For many, the natural first step to designing a data-driven rule that surpasses the appropriate levels of care and diligence in ruling out an exception is individualization: the process of tailoring a rule to the specific circumstances under consideration. In short, individualization shifts attention from the aggregate to the individual. The more individualized a rule, the more suitable it is for a particular decision-subject. For example, one way to make a data-driven rule more individualized is to add features, or inputs, to the model. A data-driven rule that uses an applicant's age, home address, and occupation in order to decide whether to grant a loan is therefore more individualized than one that uses only their age and home address.

Individualization is an information concept in that it requires a decision-maker to consider the totality of an individual's circumstances rather than make judgments based on a limited set of information. In other words, to individualize a rule is to give it additional (relevant) information. The desire for individualized decisions—the first component of our framework—is not new. Indeed, Lippert-Rasmussen (2011) discusses the right to be treated as an individual as a proposal for reducing statistical discrimination (treating an individual as if they were the statistical average of similar individuals). The push for individualization is based on the logic that, the more individualized an assessment, the less likely it is to have made broad-strokes generalizations and, as a result, to yield exceptions.

Individualization is a particularly useful concept because it appears in both legal texts (cf. the right to individualized sentencing (Berry III, 2019; Jorgensen, 2021)) as well as technical ones. As such, a law requiring individualization in data-driven rules would pave a clear path for computer scientists. Indeed, much of machine learning echoes the belief that, with enough information and enough historical data, a data-driven rule can predict the target outcome with perfect accuracy. Individualization has become so central to machine learning that data-driven rules are often justified based on their level of individualization. Most theorems in machine learning, for instance, follow the template: "As N goes to infinity, the error goes to 0" (occasionally accompanied by a "with high probability"), where N quantifies the amount of information.

Perfect individualization, however, is difficult to implement. In practice, current methods are incapable of individualizing in ways that humans do naturally. Humans, for example, are generally flexible enough to update their decisions to incorporate additional pieces of information. Although a judge may initially receive certain information about a defendant, they can update their belief when given novel information (e.g., that the defendant volunteers or has dependents). Humans rely on this unique ability to holistically examine an individual's circumstance in order to produce individualized decisions. In contrast, most (if not all) data-driven rules have fixed inputs and cannot incorporate features that are not present in the training data.

So, perhaps perfect individualization is not possible, but is individualizing the rule as much as possible (albeit imperfectly) all that is required to ensure that the rule is fit for use? Stated differently, suppose that a data-driven rule were perfectly individualized—that is, it incorporates all relevant information. Would such a fully individualized data-driven rule be enough?

3.2 Individualization is Not Enough: Uncertainty Also Matters

No—individualization is not the only pertinent factor. There are two additional components: uncertainty and harm, and we focus on the former in this section. The takeaway is that while individualizing a data-driven rule takes an important step toward ensuring that it does not neglect relevant information, no amount of individualization can remove all the uncertainty in a data-driven rule, and the amount of uncertainty matters when the risk of harm is high.

Recall that individualization is an information concept: it relies on the belief that, holding everything else equal, adding information improves a data-driven rule. Conveniently, this reasoning also underlies machine learning, which is founded on the idea that data is king (i.e., that with enough information, a data-driven rule can perform perfectly). In reality, however, even the best data-driven models make mistakes, often because some predictions are inherently impossible to get right every time. In fact, there are very few (if any) meaningful settings in which a perfect rule exists, and the main barrier is uncertainty.

To illustrate the limitations of individualization, consider the following two types of uncertainty (Kendall and Gal, 2017a):

1. *Epistemic uncertainty* is systematic or reducible uncertainty that arises from lack of knowledge. For example, a prediction of tomorrow's temperature that is based on past years' temperatures at this time of year has greater epistemic uncertainty than the prediction of tomorrow's temperature based on past years' temperature at this time of year *and* today's temperature.

2. *Aleatoric uncertainty* is statistical or irreducible uncertainty that arises from the inherent randomness or "unknowability" of an event. At the time of prediction, no information exists that can reduce this type of uncertainty. For example, the randomness in the wind patterns that may occur between today and tomorrow prevents a temperature prediction that is made today from being perfectly certain about tomorrow's temperature, and this randomness can be attributed to aleatoric uncertainty.

Through these two types of uncertainty, it becomes clear that while individualization may reduce epistemic uncertainty, it cannot reduce aleatoric uncertainty. In some cases, individualization does not even reduce epistemic uncertainty. Consider the following examples.

Example 1 (Individualization increases granularity at the risk of increasing uncertainty) *Consider a data scientist who wishes to increase the individualization of a data-driven rule used in healthcare. To do so, the data scientist adds features to the rule's input. Instead of taking in a patient's current age, height, and weight as inputs, the data scientist modifies the rule to also accept the patient's history of heights and weights at every year of their life.*

Suppose the data scientist uses a nearest-neighbors-style algorithm—an approach that makes a prediction for patient X based on previous (exemplar) patients who have similar attributes as X. Then, the more refined the features, the fewer exemplar patients for X exist. In other words, individualization reduces the amount of evidence that the nearest-neighbors rule can use to generate its assessment. As such, while the data scientist reduces epistemic uncertainty in one way, they increase it in another.

Example 2 (The unknowability of unobserved outcomes) *Consider a data-driven decision aid for college admissions—specifically, one that predicts how well a student will perform if admitted. Beyond random noise, there are multiple ways that aleatoric uncertainty arises.*

For one, even if the student is similar to previous students for which there is data, one could argue that a student's performance is not predetermined, *i.e., that they have the ability to perform differently from past individuals. That each student possesses their own potential for success—that they have their own autonomy—means that no amount of individualization can predict performance with certainty. Indeed, believing that a data-driven rule carries no uncertainty holds students responsible for the performance of previous students (namely, students in the training data). While individualization can improve a data-driven prediction, it continues to hold the decision-subject responsible for the performance of previous—albeit, increasingly similar—students, and there is always uncertainty associated with the decision-subject's own potential for success.*

For another (and perhaps more concretely), there is also omission bias. The training data only captures the performance of students who were admitted, which implies that the performance of a student who was not admitted is unknowable (Kleinberg et al., 2017). Perhaps a student who is similar to the decision-subject but was not admitted would have performed very well.

Lastly, even if a decision-maker has perfect knowledge of previous students' outcomes, any decision that is made now can only use information obtained up until this moment. There are, however, countless factors (or, in the language of causal inference, "interventions") that could influence a student's performance between the time of acceptance and graduation, such as whether they receive tutoring, who they befriend, and whether they take a part-time job. The only way that an assessment can be perfect and rid of uncertainty is for the target outcome itself to be an input to the assessment, but this logic is circular. If one could measure the target outcome, one would not need to infer it.

3.3 The Importance of Uncertainty: Weighing the Risk of Harm

In short, individualization can, at best, remove epistemic uncertainty, but no amount of individualization can remove aleatoric uncertainty. Perhaps one of the best ways to summarize this argument is via *computational irreducibility* (Wolfram, 2002). The reasoning behind this concept goes: a computer is one of many components in our world. Therefore, the complexity of a computer must be strictly lower than the complexity of the world. It follows from this logic a computer cannot predict any arbitrary outcome of interest Z (even if it was given all the historical data in the world and continually fed new data) because the complexity of the process that produces Z may be higher than the computational capacity of the computer.

That is not to say that data scientists should throw up their hands and give up. Indeed, computational irreducibility does not imply that every prediction task is hopeless. Rather, it says that uncertainty is inevitable when predicting a *complex* target outcome. However, eliminating uncertainty is besides the point. It is unreasonable to ask for a perfect data-driven rule that makes no mistakes. Instead, we ask that the level of uncertainty be *balanced against the risk of harm*.

More precisely, suppose that one of the decision outcomes would inflict significant harm. Then, no matter how individualized a decision rule may be, the decision to inflict harm should only follow if the level of certainty is high enough. If, on the other hand, the level of uncertainty (epistemic and aleatoric) is too high, then the decision-maker should err on the side of caution (less harm).

As an extreme example, suppose a decision-maker is presented with a newborn and must decide whether to confine them for the rest of their lives based on an evaluation of whether they will commit murder during their lifetime. The decision is made at the time of birth, so the only information that is available must also be available at the time of birth. A rule could be perfectly individualized (based on the information at the time of birth), but most would agree that there are so many unknowable factors that could contribute to the newborn's future actions that no amount of individualization would justify inflicting a harm as high as confining a newborn for life.

A decision outcome's risk of harm therefore determines the amount of individualization and certainty necessary to utilize a data-driven rule whose recommendation inflicts harm. Some decisions might carry a risk of harm so low the level of individualization and certainty needed to justify the use of a data-driven rule is accordingly low. It is natural to then ask: How should harm be measured? While providing an explicit framework for quantifying harm is out of the scope of this piece, we note that prior works have laid out a path for doing so, including Wachter and Mittelstadt (2019)'s work on the right to reasonable inferences (in which they discuss the determination of "high-risk inferences") and Kaminski and Urban (2021)'s right to contest AI (in which they characterize risk of harm in terms of "significant effects"). The European Union's Artificial Intelligence Act also provides a "risk methodology" for categorizing high-risk decision contexts (European Union, 2022).

3.4 Putting it All Together

Our framework requires that the decision-maker not necessarily presume that a data-driven rule is suitable for a decision-subject, particularly when the risk of harm is high. Rather, we require that the decision-maker inflict harm only if they have applied due care and due diligence in determining whether the data-driven rule is fit for use on the individual in question. It seeks to prevent data-driven decisions from inflicting irreparable and repeated harm on individuals who, through no fault of their own, are exceptions to a data-driven rule. Our work emphasizes that data-driven rules cannot be applied blanketly. While a data-driven rule may be appropriate for some individuals, it may not be appropriate for all. In particular, when a decision may inflict significant harm on the decision-subject, examining whether or not a decision-maker is justified in using the data-driven recommendation becomes pertinent. In this way, our framework in keeping with existing concepts (originally intended for human decision-makers), including the right to dignity and the right to individualized sentencing.

Importantly, our framework does not imply that data-driven rules should be dropped altogether, nor does it suggest that they be used in every case. It does not even suggest that there is a clear line between the types of decisions in which data-driven rules are appropriate (e.g., that data-driven decision aids should be used in lending but not sentencing). Rather, it argues that there are some contexts in which the stakes are so high that each decision-subject deserves appropriate consideration of whether the data-driven rule is fit for them. In the same way that certain information is discarded as irrelevant (e.g., a college admissions board may discard a student's sophomore Fall grades if a family tragedy occurred that semester), a data-driven recommendation may need to be discarded. While useful, this analogy does not carry over perfectly because it is unclear when to discard a data-driven rule. Data-driven rules behave quite differently from human ones—for instance, the "intent" and "reasoning" behind a data-driven recommendation are often inscrutable.

In this piece, we find that honoring an individual's dignity requires the consideration of three factors: individualization, uncertainty, and harm. Crucially, these three factors are not only interpretable to lawmakers, but also meaningful concepts in machine learning. They therefore provide a clear language with which to assess data-driven decisions.

More precisely, we require that the decision-maker first evaluate the level of harm of each decision outcome. Based on the level of harm, the decision-maker should then evaluate the data-driven rule based on two considerations: individualization and uncertainty. Individualization characterizes the suitability of a rule based on how much information it considers (e.g., whether it knows enough about the decision-subject or has enough training data that pertains to the decision-subject). The level of uncertainty can be divided into two types: epistemic and aleatoric. The former captures uncertainty due to lack of information, and the latter captures the inherent unknowability of a prediction task. We require that the decision-maker utilize the data-driven recommendation only if the levels of individualization and certainty are high enough to justify the level of harm that would result from that recommendation.

4 Operationalizing the Framework

In this section, we examine how this framework could be operationalized. We consider what it does (and does not) mean to invoke the framework as well as *ex ante* and *ex post* measures.

4.1 Invoking the Framework

Does invoking the framework mean proving that the data-driven rule made a mistake? Or is it that a decision-subject who does not like their decision outcome can always claim to be an exception, thus nullifying any data-driven rule in high-risk settings?

Our framework says neither. Invoking the framework is *not* equivalent to proving that the data-driven rule made a mistake. For one, the outcome of interest is not always observable. In many cases, it is impossible to determine whether a mistake was made (e.g., a judge can never know whether a defendant who is denied parole would have reoffended *if* they had been granted parole instead).

Consider the following (simplified) example. Suppose a data-driven rule delivers random recommendations. For instance, suppose that it simply flips a coin each time it is asked for a recommendation. Even this random rule is bound to be correct for some individuals. However, whether this rule happens to be correct is besides the point. If the decision's risk of harm is high (e.g., a sentencing decision), such a rule should not be applied regardless of whether or not it turns out that, down the line, the random flip happens to correctly predict the outcome. It is simply not suitable for a high-risk setting. This evaluation of a data-driven rule's suitability is what underlies our framework. Namely, the data-driven rule should only be applied if deemed suitable for the specific decision-subject, where the level of consideration must be fitting to the risk of harm (for which we provide a framework in Section 3).

Importantly (and in answer to the second question above), our work does not imply that *every* individual is an exception. That is, a decision-subject who does not like their data-driven decision outcome cannot simply reverse the decision using our framework. In fact, *a data-driven rule can still satisfy our framework even if it makes mistakes*. It is indeed unreasonable to expect a data-driven rule to never make mistakes—a decision-subject can, at best, hope that a data-driven decision-maker ensures that a data-driven recommendation is only used if it is deemed fit for the given context. Our framework captures this principle. It does not requires that a data-driven rule is perfect but that is appropriately applied.

In this way, our framework is not simply a matter of mistakes. It can be violated even when a mistake has not been made (or cannot be verified). At the same time, our framework is not necessarily violated when a mistake is made. Crucially, while a data-driven rule's accuracy—which many believe can be used to evaluate a data-driven rule's suitability (Supreme Court of Wisconsin, 2016)—is an important performance metric, it is another way of measuring mistakes. Therefore, accuracy alone cannot fully capture the suitability of a data-driven rule, as detailed in Section 3.

4.2 *Ex ante* Justification

Our framework would require an *ex ante* justification that a data-driven decision appropriately considers the three components of our work—harm, individualization, and uncertainty—before such a data-driven decision is applied. Specifically, the data-driven assessment must (1) evaluate the potential harm that the decision could inflict; (2) justify the rule on the basis of its level of individualization; and (3) demonstrate that, given the level of harm and individualization, the rule appropriately and meaningfully incorporates uncertainty into its decision *or* appropriately and meaningfully communicates it to the final decision-maker.

In order to evaluate a decision's potential harm, one can use the standard of "significant effects" in Article 22(1) of the General Data Protection Regulation (GDPR) (European Commission, 2016), as studied by Kaminski and Urban (2021). Wachter and Mittelstadt (2019) present a similar framework in their discussion of "high-risk inferences" with respect to the right to reasonable inferences. One can also turn to the European Union (2022)'s Artificial Intelligence Act, which provides a "risk methodology" for evaluating and categorizing high-risk decision contexts.

If the risk of harm is high enough, the next step is to characterize a data-driven rule's level of individualization (which may vary across decision-subjects), as given by (2). Characterizing a rule's level of individualization can be done in multiple ways. For example, one could require that a data-driven decision aid report its input variables, which reflect the data-driven rule's granularity. One could also require that the data-driven rule be evaluated on performance metrics that are more fine-grained than accuracy, such as calibration or multicalibration (Hébert-Johnson et al., 2018).

Lastly, (3) is a final step that combines the insights of (1) and (2). Specifically, (3) determines whether, given the potential harm assessed in (1) and the level of individualization found in (2), the final decision appropriately and meaningfully incorporates uncertainty. If the final decision maker is the *algorithm*, one must demonstrate that the assessment appropriately and meaningfully considers uncertainty. If the final decision-maker is *human*, then the data-driven assessment must appropriately and meaningfully communicate uncertainty to them. Incorporating uncertainty is necessary, as it synthesizes the assessments of harm and individualization. For instance, if a decision carries a risk of significant harm, then the level of individualization and the accompanying certainty may not be high enough to justify inflicting harm. It may even be the case that, in certain contexts, no matter how individualized the assessment, there is too much uncertainty to justify inflicting harm while, in others, the risk of harm is so low that a high level of uncertainty is acceptable. Meaningfully incorporating or communicating uncertainty for (3) is an active area of research in human-computer interaction (Hullman, 2016; Hofman et al., 2020). To communicate uncertainty meaningfully, the assessment could report on different types of uncertainty, similarly to how existing works distinguish between epistemic and aleatoric uncertainty (Kendall and Gal, 2017b), as discussed in Section 3.

4.3 *Ex post* Contestation

It is important that decision-subjects be able to contest a data-driven decision *ex post*. As 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.

As a possible template, one could turn to the procedure for contesting on the basis of Title VII of the U.S. Civil Rights Act's notion of disparate impact (Barocas and Selbst, 2016). Specifically, Title VII prohibits employment discrimination due to the individual's race, color, religion, sex, or national origin. A plaintiff—an individual who believes that their (potential) employer violated Title VII—can sue the employer by providing evidence of what is known as "disparate impact."

In disparate impact cases, a plaintiff must first establish that an employment practice negatively impacts a class of individuals protected by Title VII compared to its impact on individuals outside the protected class. Even if disparate impact is established, however, it can be countered if the defendant—or employer—successfully shows that the employment practice is rooted in "business necessity." The defense of "business necessity" can further be refuted if the plaintiff provides a compelling alternate employment practice that would mitigate disparate impact without violating business necessity. Contestation on the basis of our framework could mirror this three-stage procedure, as follows. First, the plaintiff must establish that (1), (2), and/or (3) from Section 4.2 has been 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 in this second stage, 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 contestation mechanisms, as surveyed by Kaminski and Urban (2021).

5 Conclusion

It is widely acknowledged that the governance of data-driven decisions requires new concepts and tools. In this work, we argue that decision-subjects are often, through no fault of their own, made to bear the burden of imperfect data-driven rules. While we cannot data-driven rules are perfect, there are several characteristics of data-driven decision-making that require special treatment. In particular, data-driven rules are not only unintuitive, but also frequently updated. As such, a persistent problem within data-driven decision-making is that it will be difficult to detect when a data-driven rule makes mistakes, imposing a high burden on decision-subjects who are made to bear the cost of being one of those individuals, i.e., of being exceptions.

In this piece, we argue that the presumption should not be that a data-driven rule—even one that has high accuracy—is suitable for an arbitrary decision-subject of interest. Rather, a decision-maker should only apply a data-driven rule if they have applied due care and due diligence (relative to the risk of harm) in excluding the possibility that the decision-subject is an exception to the given data-driven rule. In some cases, the risk of harm may be so low that only cursory consideration is required. In others, the risk of harm may be so high that a decision-maker must be convinced that the data-driven rule works well on the *specific* decision-subject of interest before applying it. We provide a three-part framework—that requires balances individualization, harm, and uncertainty—for determining whether a data-driven decision is fit for the decision-subject of interest.

References

Barocas, S. and A. D. Selbst (2016). Big Data's Disparate Impact. *California Law Review 104*, 671.

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

Citron, D. K. (2007). Technological Due Process. *Washington University Law Review 85*, 1249.

Citron, D. K. and F. Pasquale (2014). The Scored Society: Due Process for Automated Predictions. *Washington Law Review 89*, 1.

European Commission (2016). General Data Protection Regulation.

European Union (2022). The Artificial Intelligence Act.

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

Gewirth, A. (1992). Chapter 1: Human Dignity as the Basis of Rights. In M. J. Meyer and W. A. Parent (Eds.), *The Constitution of Rights*, pp. 10–28. Cornell University Press.

Glensy, R. D. (2011). The Right to Dignity. *Columbia Human Rights Law Review 43*, 1–65.

Harel, A. (2014). *Why Law Matters*. Oxford Legal Philosophy. Oxford University Press.

Hébert-Johnson, U., M. Kim, O. Reingold, and G. Rothblum (2018). Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In *International Conference on Machine Learning*, pp. 1939–1948. PMLR.

Hofman, J. M., D. G. Goldstein, and J. Hullman (2020). How Visualizing Inferential Uncertainty Can Mislead Readers About Treatment Effects in Scientific Results. In *Proceedings of the 2020 Conference on Human Factors in Computing Systems (CHI)*, pp. 1–12.

Hullman, J. (2016). Why Evaluating Uncertainty Visualization is Error Prone. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pp. 143–151.

Jorgensen, R. (2021). Algorithms and the Individual in Criminal Law. *Canadian Journal of Philosophy*, 1–17.

Kaminski, M. E. and J. M. Urban (2021). The Right to Contest AI. *Columbia Law Review 121*(7), 1957–2048.

Kant, I. (1967). *Grundlegung zur Metaphysik der Sitten*. Universal-Bibliothek. Reclam.

Kant, I. (2017). *The Metaphysics of Morals*. Cambridge Texts in the History of Philosophy. Cambridge University Press.

Kendall, A. and Y. Gal (2017a). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? *Advances in neural information processing systems 30*.

Kendall, A. and Y. Gal (2017b). 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 (Eds.), *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2017). Human Decisions and Machine Predictions. *The Quarterly Journal of Economics 133*(1), 237–293.

Lippert-Rasmussen, K. (2011). "We Are All Different": Statistical Discrimination and the Right to be Treated as an Individual. *The Journal of ethics 15*(1), 47–59.

O'Mahony, C. (2012, 03). There is no such thing as a right to dignity. *International Journal of Constitutional Law 10*(2), 551–574.

Paradis, M. (2015, December). Dignity through Law. The New Rambler.

Post, R. C. (1986). The Social Foundations of Defamation Law: Reputation and the Constitution. *California Law Review 74*, 691.

Rao, N. (2011). Three Concepts of Dignity in Constitutional Law. *Notre Dame Law Review 86*, 1–183.

Supreme Court of California (2008). In re Marriage Cases, 43 Cal.4th 757, 76 Cal. Rptr. 3d 683, 183 P.3d 384.

Supreme Court of Canada (1990). R. v. Keegstra, 3 S.C.R. 697.

Supreme Court of Wisconsin (2016). State v. Wisconsin, 881 N.W.2d 749, 754, 757.

Travis, J., B. Western, and F. S. Redburn (2014). The Growth of Incarceration in the United States: Exploring Causes and Consequences.

United Nations General Assembly (1948). Universal Declaration of Human Rights.

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

U.S. Supreme Court (1976). Woodson v. North Carolina, 428 U.S. 280, 304.

U.S. Supreme Court (1992). Planned Parenthood of Southeastern PA. v. Casey, 505 U.S. 833.

Wachter, S. and B. Mittelstadt (2019). A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI. *Columbia Business Law Review*, 494.

Wachter, S., B. Mittelstadt, and C. Russell (2021). Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai. *Computer Law & Security Review 41*, 105567.

Wolfram, S. (2002). *A New Kind of Science*. Wolfram Media.