ISSN 2281-5147

THE PERILS OF VALUE-ALIGNMENT

Robert Lee. Poe*

Abstract

This essay argues that global AI governance risks institutionalizing violations of fundamental rights. It critiques the ethical foundation of AI governance, observing that moral objectives are being prioritized over legal obligations, leading to conflicts with the rule of law. The essay calls for a re-evaluation of AI governance strategies, urging a realistic approach that respects citizens, legal precedent, and the nuanced realities of social engineering, aiming to provide an account of some of the dangers in governing artificial intelligence—with an emphasis on Justice.

Table of Contents

THE PERILS OF VALUE-ALIGNMENT 228
Abstract 228
Keywords 229
1. Introduction 229
2. Legal Design or Design made Legal? 232
3. Distributive Decisions and Consequences 235
4. Conclusion 240

* Researchet at Scuola Superiore Sant'Anna and Ph.D Candidate (National Doctorate, University of Pisa), robertlee.poe@santannapisa.it
This work is supported by the European Union's funded project Legality Attentive Data Scientists (LeADS) under Grant Agreement no. 956562

ISSN 2281-5147

Keywords

Non-discrimination – Fair Machine Learning – Distribute Justice – Legal Design– AI Ethics

1. Introduction

An alarm bell rung may be doubted by those who hear it. Neither party in that relationship is without the possibility to error in judgement due to ignorance of circumstance in the calculation of risk. We may not have all the relevant facts, and even if we did, we may not understand why that fact-set is the relevant set and not another. Written alarms seek to draw attention to a risk and should justify the attention being drawn. This alarm has been written because the field responsible for preventing algorithmic discrimination has developed the tools and methodology being used to discriminate; and AI governance, in a quest to make the world a better place, has led to the standardization of automated distributive decisions that engage in real-life, systematic violations of the fundamental right to non-discrimination (See e.g. ISO/IEC TR 24027, 2021[1] and NIST Special Publication 1270, 2022).[2]

Recent work took note of a study (Raghavan et al., 2020) which concluded that automated hiring software for pre-screening and interviewing candidates "debias" in accordance with independence-based group fairness metrics, and we argued that such practices would likely be in violation of Article 21 of the Charter of Fundamental Rights (hereafter Charter), because the Court of Justice of the European Union (Court) has found preferential treatment in the hiring context to be limited to tie-breaking scenarios (Poe & El Mestari, 2024). This is not a fringe legal conclusion, neither in the academic study of algorithmic discrimination nor in the wider EU non-discrimination law discourse. Meanwhile, human-resource software companies like Workday sell automated hiring and promotion systems throughout Europe, and they readily state in their adverts to their pursuit of a global diversity, **equity**, and inclusion policy (Global Blueprint for Belonging and Diversity, 2024).[3] Workday also readily

[1] https://www.iso.org/standard/77607.html
[2] https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf
[3] https://www.hci.org/system/files/2024-05/Global Blueprint.pdf

ISSN 2281-5147

divulges their use of the fair machine learning techniques that will likely be found to engage in unlawful discrimination (Global Impact Report on VIBE, 2024).[4]

I believe this phenomenon is the result of a number of failures, but this brief essay identifies two: a failure (1) to understand what distributive decisions are and how the process of distribution relates to the fundamental right of non-discrimination, and (2) to understand the difference between moral and legal facts and reasoning. The essay will begin with an introduction describing the risk. The first proper section of the essay will address the question of whether system design is to conform with the law or whether the law is to conform with system design when hard moral and legal questions are both relevant to design, and the second section will give a nuanced primer on distributive decisions.

First, a necessary definition: A distributive decision is a function performed by a *designer* (e.g. an authority, guardian, provider, employer) of a distributive decision-making process, allocating *distribuendum* (e.g. goods, services, burdens, offerings) among *distributees* (e.g. applicants, recipients, patients, customers, employees, citizens). This essay will rely heavily on the above definition and the terminology within it. The description is precise, but the idea is simple: a teacher distributing grades to students, an employer distributing job offers to candidates, a bank distributing loans to applicants, a state distributing benefits to citizens—all of these are distributive decisions. In other words, the study of distributive decisions then is the study of those real-life, distributive decisions that take place all throughout human activity within jurisdictions, only some of which are subject to legal scrutiny in accordance with the principle of equal treatment.

At the very least, the automated distributive decisions subject to legal scrutiny are those that have legally significant effects. But traditionally, the distributive decisions subjected to legal scrutiny are those whose authority has ". . . a direct impact on others' lives. We may be civil servants, shopkeepers, employers, landlords or doctors who decide over how public powers are used, or how private goods and services are offered. In these non-personal contexts, non-discrimination law intervenes in the choices we make . . ." (*Handbook on European Non-Discrimination Law*, 2018, p. 42) Now, of course, these are general comments about a complicated topic i.e., the scope

[4] https://www.workday.com/content/dam/web/en-us/documents/other/workday-global-impact-report.pdf

ISSN 2281-5147

of EU non-discrimination law, which takes seriously a variety of balanced interests, but the core idea is a reasonable summary based on the enumerations of Article 21 of the Charter in light of the Equality Directives and the Court's jurisprudence (e.g. in employment relations).

Fair Machine Learning (FML) is (not entirely but predominately on the policy side) the doctrinal, technical application of a philosophy known as Distributive Social Justice. *Distributive Justice* asks the question: how should the boons and burdens of a society be shared amongst the members of a society? Distributive *Social* Justice as *an* available answer to that question. Data scientists search within samples of personal and non-personal data for evidence of numerical inequalities between groups of persons in the outcomes of past distributive decisions, in an effort to determine whether those distributive decisions were Just. This approach came naturally to the field because social scientists necessarily study recorded **outcomes** of decision processes. The study of discrimination in outcomes has a long and celebrated history, and so it was naturally relied upon for the creation of the tools for addressing algorithmic discrimination in legal-technical design.

Distributive decisions are increasingly becoming automated, and the data scientist is one among others (e.g. the philosopher, legal scholar, activist, and politician) that are defining how such decisions should be made, often institutionally (limited oversight of decision-making). For the data scientist, a decision is Just if it is not biased. But bias, for the FML community, no longer means a deviation from the *true value* of a parameter or variable but instead a deviation from *group equality* or similarity in decision outcomes. This is the difference between truth as an objective and Distributive Social Justice as an objective in learning from data, described statistically—nothing fancy. Once bias is defined as group equality or similarity in decision outcomes, the data scientist deems an outcome that is unequal or dissimilar between groups of persons, Discriminatory and therefore Unjust, if the grouping represents the sub-groups of a protected characteristic (e.g. sex, gender, race, ethnicity, and so on). And when standardizing the design of automated distributive decisions, the data scientist requires distributive decisions to be Non-Discriminatory and therefore Just. That conception of Justice and Discrimination is different than what can be found in the law (I will argue for good reason), but it satisfies the moral intuitions of many (risk creation)—at least at first glance (risk reduction).

ISSN 2281-5147

This Distributive Social Justice logic is not only being applied to automated distributive decisions, but in general data quality checks. The conception is being baked into the concept of *quality* data. Information that conveys a group disparity is removed or given a threshold of acceptability. This logic has been advocated for: during the training of models, during the mining of data (automated collection process) itself (that will later train other models), during the human-verified data quality check, and during an investigation of the outcomes of automated distributive decisions. The error is subtle yet *disastrous*. Nuance is to be had and is present throughout the body of the essay and other works of the author and colleagues, but the logic of FML described above is a synthesis of the non-legal, yet highly influential logic that risks being standardized. One might feel strongly about social justice, but there is a difference between debating about the lawful scope of positive action social policy and debating about not having a scope—without scope there can be no proportionality.

2. Legal Design or Design made Legal?

First, legal *sense* needs to be distinguished from *nonsense*. In legal philosophy this is done theoretically through reliance on the separation thesis which holds that there exists a difference between what the law is and what the law ought to be. The separation thesis is well-founded given that the answer to most legal questions is straightforward due to sufficient statutory clarity, legislative history, and the past application of the law with similar fact patterns relevant to the facts in a specific case. There is neither time nor space to develop a full account of legal methodology to show the extent of its rigor here, so it will have to suffice to say that practicing lawyers can be held accountable for making frivolous legal arguments on behalf of their clients—that is, at least, in practice the formal process that distinguishes reasonable legal arguments from the unreasonable ones in accordance with professional licensing standards.

In the practice of translating fundamental rights jurisprudence into technical specifications for automated distributive decisions, the law of a given moment and jurisdiction should have priority over political or moral passion when providing legal expertise that will later guide design and the live use of such systems in a jurisdiction.

ISSN 2281-5147

A cornerstone of AI Governance generally, and Trustworthy AI specifically, is the premise that unlawfulness in AI systems inherently undermines their ethical standing—a point made repeatedly in the Guidelines for Trustworthy AI (also cited in the preamble of the AI act)—while also accommodating and encouraging the lawful, moral aims of individuals, businesses, and institutions operating within those boundaries. While many engage in and applaud going "beyond the law" to achieve a desired moral aim, the law itself must not be contradicted. And, before asking "in what direction," it should be understood that the cases which further apply, and thus define, fundamental rights are among the most controversial.

While moral conflict may sometimes be resolved through an appeal to reason, morals themselves are not the design of reason—whether natural or artificial. Morals are inherited from generation to generation, constructing our intuition through the observation of conduct and cultural learning—whether familial, local, or societal—appealing to our sense of proportionality for agreement with others, whose moral axioms are similar enough to allow for the identification of conflicts between axioms in application to a shared and relevant fact pattern. "Thought experiments" are useful for showing others that hard moral questions exist regardless of our awareness of them. But hard moral questions exist because the experimentees experience a contradiction of held axioms and force a synthesis, a moral reconstruction—prioritization via re-balancing, removal, or replacement of morals. However, there is no substitute for nuance. The moral wealth of humanity (with beautiful and tragic complexity and contradiction) cannot be expressed in a standard, nor should we delude ourselves that it might. While one's own moral compass might appear linear, non-discrimination law is emphatically not. Instead, it is proportional. The interpretation and application of the Charter of the Fundamental Rights of the European Union begins and ends at the Court of Justice of the European Union. If moral standardization is the goal, at least rely on legal standards to form the axiomatic basis.

If a moral conclusion mirrors a legal one, then at most it is a supplementary justification for lawful behavior. If a moral conclusion deviates from a legal conclusion, then it was brought about by a moral, not legal, interpretation. The difference between a hard question of morality and a hard question of law is, itself, the rule of law—through which equal treatment grew. To whatever extent the results

ISSN 2281-5147

of an ethical impact assessment have an obligatory nature, this point should be taken seriously. And, if AI ethics experts legitimize their value selection by deriving them from the fundamental rights of a given jurisdiction, then it follows that the definition should be left to fundamental rights jurisprudence. Thus, the question is not should the Court set the standard but, instead, has the standard already been set?

The question of value-alignment is not the question of whether and how an artificial intelligence can be made Good. The question of value-alignment is, in practice and in policy, whether and how current-generation AI system design (which is advancing faster than our responsible decision-making about it) can be aligned with legal obligations—both present and emerging—and their designers held accountable for violations. The objective outlined by the European Commission has been twofold: create a legal framework encouraging the adoption and innovation of AI systems in both the private and public sector, while ensuring that "high risk" purposes are brought into alignment with standards for health, safety, and the protection of fundamental rights. The enthusiastic yet cautious approach of the EU is a response to several immediately cognizable intranational and international concerns, ranging from socio-political stability to economic and defense competitiveness.

The most complex of value-alignment questions require a deep understanding, not solely of the related legal or technical subject-matter but, most importantly, the foundations which interconnect them. The problem is not that solutions are unavailable because of the novelty of artificial intelligence and the application of it. The problem is that the solutions are in the past and the researchers are in the future—more precisely in a Utopia. The crucial difference between the theoretical construction of a Utopia and the actual attempted construction of a utopia is the systematic violation of fundamental rights, always. Before the point where scholars are able to align a super-intelligent being with the Good—a dangerous and confused task to begin with—it should first be understood how to align an automated decision-making system (of any kind) with legal obligations and not the other way around. Otherwise, we are quite literally talking about a data revolution in the legal-political sense of the term.

This manuscript is about events, distributive decisions that occur in jurisdictions whose actions may have legal consequences, and if violations of equal treatment occur there may be redress for victims; but, if it is not obvious, the argument is that the

ISSN 2281-5147

definition of fundamental rights modeled technically must be modeled based on a faithful and current interpretation of the law—that means jurisdiction (cannot be global if about fundamental rights).

3. Distributive Decisions and Consequences

A few clarifications about the meaning of the term equal treatment are in order. First, *equal treatment* is a separate concept from *equal outcomes*. Outcomes may be called equal in a *factual* sense, indicating that two or more sums are identical, where those sums represent the sub-groups of a protected characteristic; in a *reactionary* sense, expressing disapproval or disgust when two or more sums are not identical, where those sums indicate differences between the sub-groups of a protected characteristic; in a *moral* and *political* sense, where theories of justice based on egalitarianism and social justice seek to resolve those differences through social policy; or in a *legal* sense, where the term is used synonymously with a term of art known as "substantive equality" which takes note of a legal fact—in this instance the Court of Justice of the European Union's jurisprudence on the application of direct and indirect discrimination doctrine and the scope of positive action social policy—and attempts to either predict how the Court might apply those legal doctrines in future cases or advising how the Court *ought* to interpret the law in those future cases in line with the goal of "substantive equality."

Treatment may be called equal in a *factual* sense, when a distributive decision does not prefer specific distributees or groups of distributees over others in relation to the creation and application of a standard. In the case of automated distributive decisions, evidence of difference in treatment in the factual sense is implemented by-design and evidence are discoverable with sufficient access to the system (more on this later). Treatment may also be called equal in a *reactionary* sense, expressing disapproval or disgust when a distributive decision does not measure each distributee or group of distributees in relation to the same standard; in a *moral* or *political* sense, where theories of justice based on merit and procedural justice seek to resolve those differences through social policy; or in a *legal* sense, where the term is used synonymously with a legal term of art known as "formal equality" which takes note of a legal fact—in this instance, again, the Court's jurisprudence on the application of direct and indirect

ISSN 2281-5147

discrimination doctrine and the scope of positive action social policies—and attempts to either predict how the Court might apply those legal doctrines in future cases or advising how the Court *ought* to interpret the law in those future cases in line with the goal of "formal equality."

Going forward I will use the term "sameness of treatment" and its derivatives for the expression of the *factual* concept of equal treatment, and I will use the term "sameness of outcome" and its derivatives to express the *factual* concept of equal outcomes. I will use the term "equal treatment" and its derivatives to express the principle of equal treatment. The principle of equal treatment can be found under Article 2 (1) of Directive 2000/78 establishing a general framework for equal treatment in employment and occupation: the principle of equal treatment "shall mean that there shall be no direct or indirect discrimination whatsoever on any of the grounds referred to in Article 1." Thus, equal treatment is satisfied where a distributive decision neither (1) prefers a sub-group *based* on a protected characteristic, or an indissociable proxy for a protected characteristic, over another in relation to the creation or application of a standard, *unless* such differential treatment can be justified by concerns of health, safety, or the protection of the rights and freedoms of others, where the means of achieving the aim are appropriate and necessary (direct discrimination); nor (2) applies a standard that disproportionately affects a sub-group of a protected characteristic *based* on a dissociable characteristic, *unless* the use of the dissociable characteristic is justified by a legitimate aim and the means of achieving that aim are appropriate and necessary (indirect discrimination).

In short, neither differences in treatment nor differences in outcome are necessarily discriminatory. I place the definitions above knowing full-well the meaning may remain elusive in the hope it becomes clearer through the explanation of distributive decisions, sameness of treatment, sameness of outcomes, equal treatment, and the principle of proportionality.

Again, a distributive decision is a function performed by a *designer* (e.g. an authority, guardian, provider) of a distributive decision-making process, allocating *distribuendum* (e.g. goods, services, burdens, offerings) among *distributees* (e.g. applicants, recipients, patients, customers, citizens). Not all distributive decisions are subject to equal treatment. But it must be remembered that distributive decisions happen in legal jurisdictions and thus are the subject of legal consideration when falling into the scope

ISSN 2281-5147

of law. A comprehensive account of legal considerations would begin with examining the "why" behind the purpose inherent in a specific distributive decision, revealing whether it falls within the scope of the law. It should then address the "who," identifying the relevant roles involved, followed by the "where," which, in part, determines the jurisdiction. Next, it should consider the "what," outlining the rights to hold and transfer the distribuendum, and conclude with the "how," detailing the means of pursuing the purpose, including infrastructural considerations for personal and non-personal data protection when distributive decisions are automated and digital or processed the old-fashioned way. Information about the design of a distributive decision can be merely stated or presumed, ex-ante conforming and evident, or, where no access to the decision-process itself is available, witnessed in the outcomes of a distributive decision. In this essay, I am concerned with the equal treatment aspect of the "how" of distributive decision-making; but in the paragraph below, I am solely concerned with the sameness of treatment aspect of the "how" of distributive decision-making.

So, how is a distributive decision made? It is made in two distinct phases: *categorical ordering* and *patterning*. **Categorical ordering** is essential to the **creation** of a standard and **patterning** is essential for the **application** of a standard. A standard is by nature a categorical ordering—distinct from patterning because it is a prerequisite of patterning—which parses the relevant from the non-relevant information for the measurement of distributees. A standard is a categorical ordering because out of all observable categories of information about distributees, through which they may be compared, it must be determined which information is useful and proper for a given measurement (feature space) (i.e. what qualities does the successful candidate have for a given job posting).

Once a standard is designed via categorical ordering, it must be designed via patterning: each distributee may be described in relation to that standard, ranked, and the distribuendum may be distributed accordingly. This process will result in inequalities in the outcomes received by distributees or groups of distributees where they are, in fact, different in relation to the standard. If a designer of a distributive decision finds different outcomes undesirable (the non-legal conception of Good or Bad Discrimination), the designer might, from the outset, create a standard which prefers certain distributees or groups of distributees over others, a re-ordering guided

ISSN 2281-5147

by a patterning dilemma: preferential purpose. For such a designer, it is necessary to not let the decision be based on a well-defined categorical ordering of correct information, because such a standard would result in different outcomes.

There are two limiting cases where sameness of treatment and sameness of outcomes are *compatible*. First, it is possible that a well-defined categorical ordering of correct information treats distributees or groups of distributees the same and, at the same time, results in the same outcomes. Such an instance is only possible where distributees or groups of distributees are, in fact, the same in measurement with the standard. Second, it is possible to create a standard of such minimal content so at to not recognize differences between distributees or groups of distributees, such that all distributees or groups of distributees are both treated the same in relation to the standard while ensuring the same outcomes. The crucial qualitative difference between the two limiting cases that ***must*** be understood is that in the first a designer reaches sameness of outcomes via factual equality (a process bounded by the categorized descriptions of distributees competing under a specific standard but cooperating in societies allowed for by such standards, i.e. the preservation of spontaneous order in centralized, distributive decisions) and, in the second, the designer reaches equal outcomes via mere presentation (a process bounded by designer choice in preference).

In between these two limiting cases, sameness of treatment and sameness of outcome are *incompatible*. A designer may set a threshold at any point between these two extremes, demarcating how much difference in outcomes between distributees or groups of distributees is acceptable in their distributive decision, and by doing so, necessarily demarcates how much difference in treatment between distributees or groups of distributes is acceptable. Properly understood, the quantitative trade-off of the threshold is the sameness of treatment in one hand of a designer and the sameness of outcomes in the other.

Example A: imagine an automated loan approval system, where the algorithm is designed to approve loans based on credit score (categorical ordering: e.g. on-time payment history, credit-debt ratio, open and closed lines of credit, and so on, to predict the likelihood of loan default), it may result in different approval rates among various sub-groups of protected characteristics, not because it is *based* (if protected characteristic or indissociable proxy was not a category) on those protected

ISSN 2281-5147

characteristics but because there may be a correlation between the protected characteristic and creditworthiness in the target population. By setting a threshold (necessary use of protected characteristic or an indissociable proxy) for acceptable variance in approval rates between sub-groups of a protected characteristic (sameness of outcomes, difference in treatment), the bank increases the likelihood that applicants who would have been deemed uncreditworthy for the loan specifics are given loans—putting the borrower at risk of loan default by-default (based on the protected characteristic) and the lender at risk of a loss. Remember, while a bank is likely to shift that burden to others via fraud and/or taxpayer bailouts or socialization, the borrower will be left with little recourse.

Example B: imagine an automated hiring system, where the algorithm is designed to evaluate candidates based on qualifications such as education, work experience, and skill assessments (categorical ordering: e.g., degree level, years of relevant experience, results of standardized tests, and so on, to predict job performance). It may result in different hiring rates among various sub-groups of protected characteristics, not because it is based (if protected characteristic or indissociable proxy was not a category) on those protected characteristics but because there may be a correlation between the protected characteristic and certain qualifications in the target population. By setting a threshold (necessary use of protected characteristic or an indissociable proxy) for acceptable variance in hiring rates between sub-groups of a protected characteristic (sameness of outcomes, difference in treatment), the company increases the likelihood that candidates who would have been deemed less qualified for the job specifics are hired—putting the company at risk of decreased performance and the candidate at risk of struggling in the role. Remember, while a company may try to address performance gaps through training or reassigning roles, the hired candidate might face challenges in job satisfaction and career growth.

When a distributive decision process is automated, evidence of difference in treatment exists. In short, it exists in the space between a representative sample and generalizable hypothesis assumptions, and so by measuring the relationships between distributees or groups of distributees in the data sample and comparing them with the relationships in the outcomes of an automated distributive decision, violations of sameness of treatment are quantitatively detectable—relevant to direct discrimination in the same way differences in outcome are relevant to indirect discrimination.

ISSN 2281-5147

4. Conclusion

As we move forward in the development and standardization of AI systems, it is crucial not to oversimplify the complex issues at hand. The questions surrounding AI governance and the alignment of technology with legal standards are far from straightforward. This essay offers just a glimpse into the intricate challenges we face—challenges that require deep, ongoing analysis and public discourse. While I have highlighted some critical points, this discussion is only a piece of a much larger puzzle.

The development of AI systems cannot be separated from the legal frameworks that protect fundamental rights. As we standardize technologies that have the potential to profoundly impact society, we must ensure that these standards are informed by a nuanced understanding of both legal obligations and the realities of social engineering. The issues at stake are not just technical or academic; they have real consequences for the lives of individuals and the functioning of societies. Therefore, it is imperative that these debates extend beyond closed circles of experts and become part of a broader public discourse.

ISSN 2281-5147