# "Ideal Theory" as Ideology

CHARLES W. MILLS

Starts out with background on how they came were inspired to develop non-ideal theories toward ethical study:

- Feminist ethics has evolved beyond focusing on "care" to encompass diverse perspectives, connecting with mainstream moral theory → common denominator is the goal of understanding and ending women's oppression.

The paper highlights the distinction between idealizing and non-idealizing approaches to ethical theory, with the latter being more developed in feminist theory.

- Non-idealizing approaches can potentially address the concerns of various oppressed groups, including men and women, can avoid particularism and relativism.
- Non-idealizing approaches can engage with mainstream ethics on its own terms, making it harder to ignore and marginalize.
- The dominant ideal theory in mainstream ethics can be obfuscatory and ideological, perpetuating group privilege.
- The best way to realize the ideal is through the recognition of the importance of theorizing the *nonideal.*

<u>*Vices of Ideal Theory:*</u>

- Differentiate between various senses of "ideal":
    - *Ideal-as-normative*: since ethics deals w/ normative and not factual issues, it involves the appeal to values; uncontroversial normative sense of "ideal"
    - *Ideal-as-model*: the central focus of the piece, referring to a representation of a phenomenon (P).
        - *Ideal-as-descriptive-model*: a model that aims to describe P's crucial aspects and how it works.
        - *Ideal-as-idealized-model*: an exemplar of what an ideal P should be like.
- Ideal-as-idealized-model is only useful if sufficiently close to descriptive → When the actual P is significantly different from the ideal P, it is important to work with the nonideal, ideal-as-descriptive-model, to understand what prevents it from attaining ideality.
- *Key features of ideal theory*:

    - Focuses on perfect scenarios and people, rather than real-world situations. For example, it assumes everyone always acts morally and has ideal human abilities.
    - <u>Idealized social ontology</u>: it simplifies the view of people, ignoring power imbalances and social hierarchies.
    - <u>Idealized capacities</u>: assumes people have unrealistic abilities, not considering how inequality affects development.
    - <u>Silence on oppression</u>: doesn't address historical or ongoing discrimination and its effects on society.
    - <u>Ideal social institutions</u>: assumes institutions like the family, economy, and legal system are fair and just for everyone.
    - <u>Idealized cognitive sphere:</u> presumes clear and unbiased understanding of society, neglecting the role of ideologies and personal experiences.
    - <u>Strict compliance</u>: everyone acts justly and supports just institutions, as seen in John Rawls' "A Theory of Justice."

- Marginalized groups, like feminists, Marxists, and racially oppressed people, have long been drawn to nonideal or naturalized theories → emphasize the need for theories that address real-world, non-ideal conditions, reflect their experiences.
- *Arguments for ideal theory?*
    - Moral theory not concerned with realistic assumptions about humans: doesn't work, as ethics must consider what humans can achieve
    - Moral theory is only concerned with mapping beautiful ideals, not their implementation: abandonment of the traditional goals of ethics to link to practical reason.
- *Why is it dominant?*
    - Reflects the interests and experiences of a privileged minority, such as middle-to-upper-class white males, who are over-represented in the field.
    - Privileged group's experience of reality aligns more closely with the ideal, resulting in less cognitive dissonance.
    - Does not serve the interests of marginalized groups, such as women, people of color, and the working class.

<u>*Virtues of Non-Ideal Theory:*</u>

- Lack of fit between dominant abstractions and the experiences of marginalized groups is due to the abstractions being of the ideal-as-idealized-model kind, rather than the ideal-as-descriptive-model kind.
- *Generalism v/s particularism*:

    - Particularism, whether based on individual or group experiences, has several limitations and dangers, as it can lead to intellectual isolation, relativism, and an inability to effectively challenge hegemonic understandings of societal issues.
    - Issue with generalism is that it abstracts from ideal-as-idealized-models not ideal-as-descriptive-model, that capture the essentials of the situation of women and nonwhites.
    - Author advocates for generalism/abstraction from ideal-as-descriptive-models , which reflect experiences of marginalized groups while avoiding the pitfalls of particularism.

*Non-idealized descriptive mapping concepts:*

- Moral cognition == perception+concepts
- But all “theorizing takes place in an intellectual realm dominated by concepts, assumptions, norms, values, and framing perspectives that reflect the experience and group interests of the privileged group (whether the bourgeoisie, or men, or white)”
- Standpoint theory: certain realities are more visible from the perspective of the subordinated than the privileged; crucial conceptual innovation often comes from marginalized groups.
- Nonideal theory acknowledges that people's cognitive abilities are influenced by their social location, and therefore strives to challenge and reconstruct dominant conceptual tools and boundaries to better understand oppression and its consequences.

*Normative Concepts:*

- Critiques the adequacy of ideal theory in dealing with normative concepts
- Argues that its inherent limitations can lead to the perpetuation of harmful ideals and the overlooking of important perspectives and experiences from socially subordinated groups.
- Advocates for a nonideal theory that is more sensitive to the realities of oppression and better equipped to challenge and reconstruct dominant conceptual frameworks.

*Non-idealized theory already contained in Ideal theory?:*

- Argues against the claim that nonideal theory is contained within ideal theory.
- Claim fails to recognize the cognitive challenges and obstacles involved in rethinking dominant paradigms and misconceptions.
- Extending ideal theory to include marginalized groups is not straightforward; if it were that simple, historically prominent philosophers would not have excluded women and nonwhite individuals from their theories.

- Feminist critiques show that including women requires rethinking how equal rights and freedoms work in the context of female subordination, rather than merely extending existing concepts to them. This requires empirical input and a focus on nonideal realities, which separates it from ideal theory.

*Final Thoughts*

- Nonideal theory is better at realizing ideals by acknowledging and addressing the obstacles that hinder their acceptance and implementation.
- The debate between ideal and nonideal theory reflects a larger philosophical conflict between idealism and materialism.
- Materialism, in this context, refers to grounding moral theory in society and recognizing the influence of social structures and privileges.
- Understanding how social location can blind individuals to important realities and maintain the status quo is crucial for initiating social change.
- Ideal theory often disregards these problems or assumes that better arguments alone can solve them.
- Ideal can be achieved more effectively by acknowledging the nonideal, and assuming theideal without addressing the nonideal only perpetuates existing issues.

# Inherent Trade-Offs in the Fair Determination of Risk Scores (Kleinberg et al. 2016)

- Relevant background context
    - ProPublica Machine Bias piece: analysis of Northpointe's COMPAS recidivism algorithm highlighting how Black individuals were assigned higher risk scores than white individuals
    - Northpointe responded by countering that their algorithm is calibrated
        - "Northpointe unequivocally rejects the ProPublica conclusion of racial bias in the COMPAS risk scales."
        - "ProPublica focused on classification statistics that did not take into account the different base rates of recidivism for blacks and whites. Their use of these statistics resulted in false assertions in their article that were repeated subsequently in interviews and in articles in the national media."
        - "When the correct classification statistics are used, the data does not substantiate the ProPublica claim of racial bias towards blacks."
    - Kleinberg et al. (in this piece) and Chouldechova (in Fair Prediction with Disparate Impact) highlight the impossibility of simultaneously satisfying calibration and fairness except in ideal scenarios
- Desirable properties for fairness:
    - Calibration: for any population subgroup, if members in a subset of the subgroup are predicted as having a risk score of x%, then x% of those members should actually have positive ground-truth
        - Concrete example: if a recidivism algorithm predicts a subset of women of having a risk score of 70% and is well-calibrated, then 70% of those women should actually reoffend
        - Intuition: "scores mean what they claim to mean, even when considered separately in each group"; allows for comparing individuals across groups with the same risk scores (otherwise would have to consider things like base rates, societal biases, etc.)
    - Negative class balance: average risk score given negative ground-truth and membership of one group should equal average risk score given negative ground-truth and membership of a different group
        - Concrete example: average risk score given negative ground-truth is 30% regardless of race
    - Positive class balance: same as above (by symmetry) except that average risk scores given positive ground-truth should be equal
        - Intuition: individuals with similar behavior across groups should be treated similarly
- Analysis of the COMPAS recidivism algorithm revealed that it was indeed calibrated but failed the positive/negative balance conditions
- Kleinberg et al. demonstrate that their three conditions can only be simultaneously achieved whenever base rates are equal or risk prediction is perfect

    - These conditions should not be expected to hold in practice
- Lots of math and relaxations of the original problems and theorems follow…
    - Main result with epsilon approximation
    - Exploration of designing well-calibrated systems that are mostly fair
        - Involves reduction of integral fair risk assignments to NP-complete problem
- Conclusion
    - Simple base rate differences or predictive inaccuracies will lead to unsatisfying consequences in terms of predictive differences across groups
    - Finding a solution that best navigates these tradeoffs may be computationally infeasible
    - Figuring out conditions under which 2 of 3 conditions can be satisfied in the case that false positives and false negatives are associated with different costs may be room for future work
- Epilogue
    - ProPublica described their analysis methodology in another post and released their data on GitHub
        - Resulting data is effectively the MNIST of the FAccT world
    - Northpointe got some rebranding
        - “On January 9, 2017, Courtview Justice Solutions Inc., Constellation Justice Systems Inc., and Northpointe Inc. were united as equivant under a single brand dedicated to helping justice agencies better serve our communities. Our mission is to embrace community while advancing justice, delivering better outcomes for all who touch the justice system.”

# Handout for "Algorithmic Fairness from a Non-Ideal Perspective"

Conny Knieling, April 26 2023

## 1 - Overall Approach

| Political Philosophy | Algorithmic Fairness |
|---|---|
| - Justice, The Good; Large scale | - Fair Algorithms; Small-Scale/Single Task assessing and managing disparities among groups in connection with the deployment of ML-supported decision systems |
| - Target: Injustice (in Institutions, Societies, …) | - Target: Unfair, Biased Algorithms |
| - Offering practical guidance | - Offering single solutions |
| - Multiple decision-makers | - Individual decision-maker |
| Ideal vs. non-ideal theory | |

**MAIN THESIS: How a common distinction in political philosophy can help address issues in the fair ML literature**

- How this distinction and a different approach can be put to work in understanding and addressing algorithmic injustice
- This distinction and in particular non-ideal theorizing in political philosophy provides a fruitful lens for formulating strategies for addressing algorithmic injustice
- Injustice instead of Fairness
- We run into issues that are similar (or the same) as in ideal theory approaches in political philosophy

**Question:**

Is the distinction described correctly?
Can we transfer/apply this in ML?
Is this correctly transferred?
Is the transfer helpful?

## 2 - Ideal vs. Non-ideal methodological approaches

- Useful distinction in political philosophy
- First proposed by John Rawls in *A Theory of Justice*:
  ideal theory provides a necessary base for non-ideal theories to be built on it
- Refers to ways of understanding the demands of justice at large,
  And offering practical normative guidance to institutions for complying with these demands
- Ideal theory heavily criticized by e.g. Mills
- Different ideas about what this distinction is or should be (e.g. Laura Valentini)

### Ideal approach:

Start by articulating a conception of an ideally just world under a set of idealized conditions
Two functions:

1. Provides decision-makers with a target to aspire towards
2. Serves as an evaluative standard for identifying and assessing current injustices; closing gap between ideal and reality

Architecture-Approach, planning or setting out an ideal world to then strive for it:

## Non-ideal approach:

Emerged as a result of challenges to ideal modes of theorizing.
Three challenges that motivate this approach:

1. "When we consider **the intended role** of a conception of an ideally just world for diagnosing actual injustice"
   Issue is: static nature of ideal state, pursuant diagnostic lens; identifying injustice in terms of difference between actual and the ideally-just world
   - The conceptual framing might be wrong, leaving out injustices
   - Might overlook factors that gave rise to injustice in the first place
   - Causal Factors: Historical origins, dynamics of current injustice, ongoing social forces that sustain them are absent
   - These evaluative standards distort our understanding of current injustices

   ➢ Inefficient, inadequate, and unjust (ignoring causes)?

2. "Employing a conception of an ideally just world as an evaluative standard is **not sufficient** for deciding how actual injustices should be mitigated"
   - Any difference might be interpreted naively as a cause of an injustice
     → Not only ineffective but also could be exacerbate the problem they try to fix
     But actual world can deviate in many ways, same difference can have different causes
   - Any policy aiming to close that gap might be naively seen as justice-promoting
   - Simultaneously eliminating all discrepancies can be impossible
     → Need guidance for determining which gaps matter

   Example: just society as race-blind, solution to injustice: end race-conscious policies (unjust, inefficient)

   ➢ Inefficient, not sufficient, potentially dangerous and unjust (ignoring causes)?

3. "Concerns the **practical usefulness** of the ideal approach for current decision-makers"
   - Given the idealized assumptions under which ideal theorizing proceeds, the use-ability of the ideal approach seems hard to impossible
   - fails to answer the questions of what we might reasonably expect from a decision-maker in the real world

   Example: strict compliance (frequently assumed by ideal theorists): all agents comply with what justice demands of them; idealizes away real situations; is not useful for decision-makers

   ➢ Useless, not sufficient

**In summary, ideal modes of theorizing:**
(1) can lead to systematic neglects of some injustices and distort our understanding of other injustices
(2) do not, by themselves, offer sufficient practical guidance about what should be done; can lead to misguidance
(3) do not, by themselves, make clear who, among decision-makers, is responsible for intervening to right specific injustices

## Non-ideal approach, defined:

Begins by identifying actual injustices that are of concern to decision-makers or are raised by those who are affected
Troubleshooting-approach, addressing actual concerns and complaints:

> Anderson: "[Non-ideal theorists] seek a causal explanation of the problem to determine what can and ought to be done about it, and who should be charged with correcting it. This requires an evaluation of the mechanisms causing the problem, as well as responsibilities of different agents to alter these mechanisms" [1, p. 22]

Still: crucial role for normative ideals within this approach, but importantly different:

- Normative ideals have different roles than the role of ideals in ideal approach
    - Normative ideals in ideal theory are extra-empirical, i.e. they set the evaluative standard against which actual practices are assessed without themselves being empirically evaluated
    - Normative ideals in non-ideal theory act as **hypotheses** about potential solutions; subject to revision in light of their efficacy

# 3 - Algorithmic Fairness

Most works on algorithmic fairness have the more restricted aim of assessing and managing various disparities that arise among demographic groups in connection with the deployment of ML-supported decision systems

Smaller scale, Individual decision-makers instead of multiple

**Argument for**: the dominant approach in algorithmic fairness can be seen as exercises in small-scale ideal theorizing

## Fairness as ideal theorizing:

1. **Fairness Ideal:**
   Typically researchers begin by outlining a conception of a "fairness ideal"
   Often place fair ideal at the group level (e.g. race, ethnic origin, sex, and religion - protected groups)

   Examples:
    - impact different protected groups in the same way
    - membership in a protected group is irrelevant or does not make a difference to the allocative procedure
    - a treatment disparity might exist in a fair state, if it is justified by the legitimate aims of the distributive procedure

   Based on: historical legal cases, people's intuitive judgments, works in political philosophy

2. **Fairness Metric:**
   Next, on the basis of this ideal, researchers specify a quantitative evaluative standard for diagnosing potential allocative injustices and guiding mitigation efforts
   Often: mathematical expressions that quantify how far two of the protected groups are from parity

Metric: measuring the magnitude of (dis)parity, formal proxy for the degree of divergence from the ideal

Fairness ideals do not fully determine the metric, further value judgments can be necessary

Examples:

- The ideal that membership in protected groups should be irrelevant to allocative decisions can be articulated in the language of statistics by requiring the outcome be independent (probabilistically) of the protected attributes. However, the same ideal can also be expressed in the language of causality, e.g., by requiring that the average causal effect of protected attributes be negligible.

Even when being motivated by the same ideal, such fairness metrics make different demands from the user and can result in different verdicts about the same case:
causal metrics additionally require the acquisition of a causal model that faithfully describes the data-generating processes and for which the desired causal effect is identifiable. This can be impossible in some cases!

3. **Justice as Deviation:**
current approaches seek to promote fairness (or mitigate unfairness) by modifying ML algorithms to maximize utility subject to a parity constraint expressed in terms of the proposed fairness metric

fairness-enforcing modifications can take the form of interventions:
(i) in the pre-processing stage to produce “fair representations”
(ii) in the learning stage to create “fair learning”; or
(iii) in the post-processing by adjusting the decision thresholds
In all cases, the range of solutions to algorithmic harms is limited to an intervention to the ML algorithm.

Absent is the broader context in which the “certifiably fair” model will be deployed.
Recalling Anderson’s critique of ideal approaches,

- neither the mechanisms causing the problem,
- nor the consequences of algorithmically-guided decisions,
- nor “the responsibilities of different agents to alter these mechanisms” are captured.

## 4 - Troubles with Ideal Fairness Metrics

If: current works on algorithmic fairness pursue (small-scale) ideal theorizing,
Then: we should expect these works to encounter the same types of challenges as those confronting ideal theorizing more generally.

As explained above, according to critics, ideal modes of theorizing
(1) systematically neglects of some injustices; and distort our understanding of other injustices.
(2) does not offer sufficient practical guidance and can lead to misguided mitigation strategies.
(3) fails to delineate the responsibilities of current decision-makers in a world where others fail to comply with their responsibilities.

Now, address each of these challenges in turn, and show that these same types of worries arise with respect to current works on algorithmic fairness:

### Systematic Neglects of Rights

- Identification of injustices in ideal theorizing is constrained by the underlying conceptual framing of normative ideals
If this conceptual framing is not sufficiently rich or comprehensive, we run the risk of overlooking many actual injustices.

- Virtually all algorithmic fairness ideals are framed in comparative terms.

- But: In some cases an individual's just due is determinable independent of any comparison and solely by reference to how that individual should have been treated in light of her rights and deserts. There are comparative as well as non-comparative cases of injustice
    - Example: Human Rights Violations of everyone

- Troubling even with respect to comparative cases of injustice:
  Due to their narrow focus, fairness metrics essentially take the set of protected classes to exhaust comparison classes that might matter from the perspective of justice and fairness. The complete reliance of such metrics on the particular specification of relevant comparison groups limits their adequacy in this regard.
    - Example: Employee-Scenario where height and weight become causal factors

- ? Unconstrained by these demands of comparative justice, algorithmic-based decisions might result in the creation of new "protected groups"

## Distortion of the Harms of Discrimination

- Any divergence from the ideal of parity among protected classes (potentially subject to certain qualifications) is identified as a case of unfairness

- fairness metrics based on these ideals often have the property of being anonymous or symmetric; whether a distribution of benefits and burdens is fair does not depend on who the affected individuals or groups are.

- anonymity is not always a desirable property
- evaluating fairness claims requires going beyond the observation that some disparity exists

- We need to know why the disparity exists and to understand "the processes that produce or maintain it"

  Example: Admission Differences

  Approaches that incorporate knowledge of demographic labels are colloquially referred to as "fairness through awareness"
  However, awareness of demographic membership alone is too shallow to distinguish between situations

- Require a deeper awareness, not only of demographic membership but of the societal mechanisms that charge them with social significance in the given context and that give rise to existing disparities.

- especially problematic for statistical metrics but not much better for recently-proposed causal approaches,

  similarly insufficient for capturing when a given disparity is reflective of discrimination, let alone whose discrimination it might reflect or providing guidance as to when the current decision-maker has a responsibility or license to intervene.
  Also,
    - Relationship between causal mechanism and responsibilities of the current decision-maker?
    - accounting of the causal mechanisms?

## Insufficient Insights and Misguided Mitigation

- Insofar as the underlying causes of preexisting disparities and the consequences of proposed policies are ignored, these mitigation techniques might have adverse effects

  Example: disparate learning processes (DLPs), jointly satisfy two parities, blindness and demographic parity
  However, DLPs are oblivious to the underlying causal mechanisms of potential disparities and in some cases, DLPs achieve parity between protected classes (e.g., genders) by giving weight to the irrelevant proxies (e.g. hair length)

  DLP risk amplifying the very injustices it is intended to address
  They can reinforce social stereotypes

- neglect considerations about whether the enforced parity might in fact result in long term harms

### Lack of Practical Guidance

- Current approaches to algorithmic fairness seek to address “is there discrimination?” but leave open the questions of “who discriminated?” and “what are the responsibilities of the current decision-maker?”

- identifying statistical disparities may be valuable unto itself, e.g., as a first step to indicate particular situations that warrant investigation, it provides little moral or legal guidance to the decision-maker

- providing normative guidance requires identifying not only what would constitute a just world but also what constitute just decisions in the actual world, with its history of injustice

## 5 - Discussion

- have not failed to deliver in the marketplace
  From the perspective of stakeholders, these metrics offer an alluring solution: continue to deploy machine learning systems per the status quo, but use some chosen parity metric to claim a certificate of fairness
  (a shield against criticism)

- In socially-consequential settings, requiring caution or even abstention (from applying ML) such as criminal justice and hiring, fair ML offers an apparent academic stamp of approval.

- literature on fair machine learning bears some responsibility for this state of affairs

? - Lacking the basic primitives required to make the relevant moral distinctions, when blindly optimized, these metrics are as likely to cause harm as to mitigate it. Thus run the risk of serving as solutionism instead of solutions

### Re-interpreting Impossibility Results

- gives a new perspective for parsing the numerous impossibility results:
  sometimes misinterpreted as communicating that fairness is impossible. However, through the non-ideal lens, these impossibility theorems are simply a frank confirmation of the fact that we do not live in an ideal world

- indicates only that our present decision-maker cannot through their actions alone bring about the immediate end to all disparity;
  ? Apply to philosophy: make evident an often overlooked shortcoming with the ideal approach

### Towards a Non-Ideal Perspective

what precisely a non-ideal approach might look like in practice

To begin, non-ideal theorizing about the demands of justice is a fact-sensitive exercise. Offering normative prescriptions to guide actions requires understanding the relevant causal mechanisms that (i) account for present injustices; and (ii) govern the impact of proposed interventions.

#### Empirical understanding of the problem

- coupled ethical-epistemic approach: not only of ethical import; it also has important epistemic implications, as it shapes how attributes should be understood and what causal relation it might bear to other factors of interest

#### Empirically-informed choice of treatment

- Deployment of predictive models constitutes a social intervention

- So far, the evaluation of these interventions is local and static

- In contrast, a non-ideal approach to offering normative guidance should be based on evaluating the situated and system-wide (involving not just the predictive model but also the broader social context, actors, and users) and dynamic (evolving over longer periods) impact of potential fairness-promoting interventions

- the appropriate judgments simply cannot be made based on the reductive description upon which most statistical fair ML operates. Developing a coherent non-ideal approach requires (for the foreseeable future) human thought, both to understand the social context and to make the relevant normative judgments.

## 6 - Conclusion

Approaching the issue of algorithmic fairness from a non-ideal perspective requires:

- a **broadening of scope beyond parity-constrained predictive models**
- **considering the wider socio-technological system** consisting of human users, who informed by these models, make decisions in particular contexts and towards particular aims.

Effectively addressing algorithmic harms demands nothing short of this **broader, human-centered perspective**, as it enables the formulation of novel and potentially more effective mitigation strategies that are not restricted to simple modifications of existing ML algorithms.