Fairness in Machine Learning and AI

Romain Couillet

romain.couillet@gipsa-lab.grenoble-inp.fr

GIPSA-lab, University Grenoble-Alps

March 24, 2021

Outline

When machines replace humans

The new era of machine learning:

When machines replace humans

The new era of machine learning:

- algorithms and machines take increasingly more decisions influencing society

When machines replace humans

The new era of machine learning:

- algorithms and machines take increasingly more decisions influencing society

These decisions:

When machines replace humans

The new era of machine learning:

- algorithms and machines take increasingly more decisions influencing society

These decisions:

- involve human (and all living) beings more or less directly

When machines replace humans

The new era of machine learning:

- algorithms and machines take increasingly more decisions influencing society

These decisions:

- involve human (and all living) beings more or less directly
- enter the realms of law and ethics

When machines replace humans

The new era of machine learning:

- algorithms and machines take increasingly more decisions influencing society

These decisions:

- involve human (and all living) beings more or less directly
- enter the realms of law and ethics

Law, ethics, and machines:

When machines replace humans

The new era of machine learning:

- algorithms and machines take increasingly more decisions influencing society

These decisions:

- involve human (and all living) beings more or less directly
- enter the realms of law and ethics

Law, ethics, and machines:

- machines have no legal identity,

When machines replace humans

The new era of machine learning:

- algorithms and machines take increasingly more decisions influencing society

These decisions:

- involve human (and all living) beings more or less directly
- enter the realms of law and ethics

Law, ethics, and machines:

- machines have no legal identity,no legal responsibility

When machines replace humans

The new era of machine learning:

- algorithms and machines take increasingly more decisions influencing society

These decisions:

- involve human (and all living) beings more or less directly
- enter the realms of law and ethics

Law, ethics, and machines:

- machines have no legal identity,no legal responsibility
- this creates many loopholes in present law terms

When machines replace humans

The new era of machine learning:

- algorithms and machines take increasingly more decisions influencing society

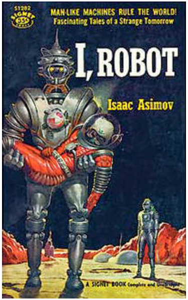
These decisions:

- involve human (and all living) beings more or less directly
- enter the realms of law and ethics

Law, ethics, and machines:

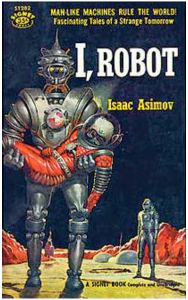
- machines have no legal identity,no legal responsibility
- this creates many loopholes in present law terms (example of self-driving cars involved in accidents!)

Isaac Asimov's three laws of robotics



I, Robot: In the 1950's, Asimov prophesied the need for laws to rule robots and machines

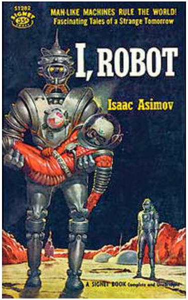
Isaac Asimov's three laws of robotics



I, Robot: In the 1950's, Asimov prophesied the need for laws to rule robots and machines

1. **First Law**: "A robot may not injure a human being or, through inaction, allow a human being to come to harm."

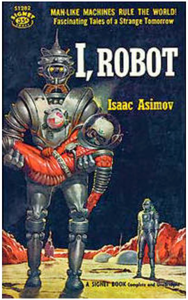
Isaac Asimov's three laws of robotics



I, Robot: In the 1950's, Asimov prophesied the need for laws to rule robots and machines

1. **First Law**: "A robot may not injure a human being or, through inaction, allow a human being to come to harm."
2. **Second Law**: "A robot must obey the orders given it by human beings except where such orders would conflict with the First Law."

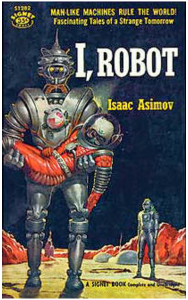
Isaac Asimov's three laws of robotics



I, Robot: In the 1950's, Asimov prophesied the need for laws to rule robots and machines

1. **First Law**: "A robot may not injure a human being or, through inaction, allow a human being to come to harm."
2. **Second Law**: "A robot must obey the orders given it by human beings except where such orders would conflict with the First Law."
3. **Third Law**: "A robot must protect its own existence as long as such protection does not conflict with the First or Second Law."

Isaac Asimov's three laws of robotics

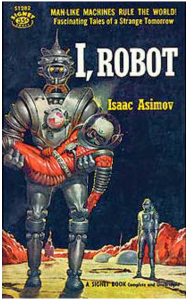


I, Robot: In the 1950's, Asimov prophesied the need for laws to rule robots and machines

1. **First Law**: "A robot may not injure a human being or, through inaction, allow a human being to come to harm."
2. **Second Law**: "A robot must obey the orders given it by human beings except where such orders would conflict with the First Law."
3. **Third Law**: "A robot must protect its own existence as long as such protection does not conflict with the First or Second Law."

⇒ Based on 3 desiderata, Asimov wrote many books on the topics, where **robots seemingly do not abide by the laws...**

Isaac Asimov's three laws of robotics



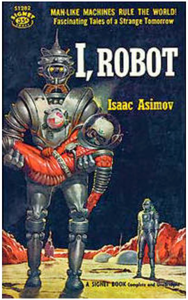
I, Robot: In the 1950's, Asimov prophesied the need for laws to rule robots and machines

1. **First Law**: "A robot may not injure a human being or, through inaction, allow a human being to come to harm."
2. **Second Law**: "A robot must obey the orders given it by human beings except where such orders would conflict with the First Law."
3. **Third Law**: "A robot must protect its own existence as long as such protection does not conflict with the First or Second Law."

⇒ Based on 3 desiderata, Asimov wrote many books on the topics, where **robots seemingly do not abide by the laws...**

Not so far from present class!

Isaac Asimov's three laws of robotics



I, Robot: In the 1950's, Asimov prophesied the need for laws to rule robots and machines

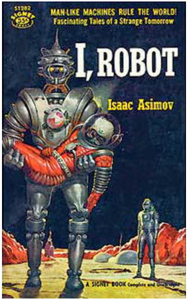
1. **First Law**: "A robot may not injure a human being or, through inaction, allow a human being to come to harm."
2. **Second Law**: "A robot must obey the orders given it by human beings except where such orders would conflict with the First Law."
3. **Third Law**: "A robot must protect its own existence as long as such protection does not conflict with the First or Second Law."

$\Rightarrow$ Based on 3 desiderata, Asimov wrote many books on the topics, where **robots seemingly do not abide by the laws...**

Not so far from present class!

- we will also meet 3 desiderata for fairness in AI "robots"

Isaac Asimov's three laws of robotics



I, Robot: In the 1950's, Asimov prophesied the need for laws to rule robots and machines

1. **First Law**: "A robot may not injure a human being or, through inaction, allow a human being to come to harm."
2. **Second Law**: "A robot must obey the orders given it by human beings except where such orders would conflict with the First Law."
3. **Third Law**: "A robot must protect its own existence as long as such protection does not conflict with the First or Second Law."

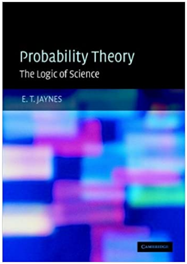
$\Rightarrow$ Based on 3 desiderata, Asimov wrote many books on the topics, where **robots seemingly do not abide by the laws...**

Not so far from present class!

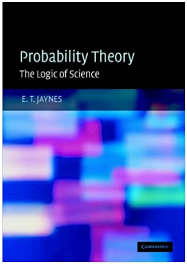
- we will also meet 3 desiderata for fairness in AI "robots"
- these will fail to be satisfying as mutually incompatible (unless in trivial cases)

From SciFi to maths: Jaynes' probability as extended logic

Probability Theory: the Logic of Science: In 2003, Jaynes **theorizes plausible reasoning**



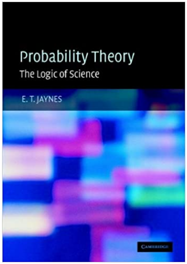
From SciFi to maths: Jaynes' probability as extended logic



Probability Theory: the Logic of Science: In 2003, Jaynes **theorizes plausible reasoning**

1. **Desideratum 1**: "Plausibility is represented by continuous real numbers."

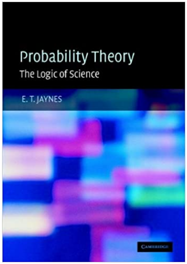
From SciFi to maths: Jaynes' probability as extended logic



Probability Theory: the Logic of Science: In 2003, Jaynes **theorizes plausible reasoning**

1. **Desideratum 1**: "Plausibility is represented by continuous real numbers."
2. **Desideratum 2**: "Qualitative correspondence with common sense." (**compatible with binary logic**)

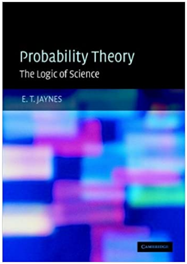
From SciFi to maths: Jaynes' probability as extended logic



Probability Theory: the Logic of Science: In 2003, Jaynes **theorizes plausible reasoning**

1. **Desideratum 1**: "Plausibility is represented by continuous real numbers."
2. **Desideratum 2**: "Qualitative correspondence with common sense." (**compatible with binary logic**)
3. **Desideratum 3**: "Consistent reasoning."
   - *Path independence:* If an answer can be calculated many ways, each should give the same answer.
   - *Non-ideological:* The reasoner does not leave out information.
   - *Equivalence:* Equivalent states of knowledge are represented by the same number.

From SciFi to maths: Jaynes' probability as extended logic

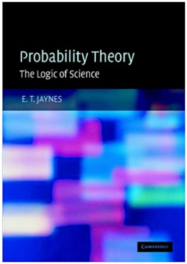


Probability Theory: the Logic of Science: In 2003, Jaynes **theorizes plausible reasoning**

1. **Desideratum 1**: "Plausibility is represented by continuous real numbers."
2. **Desideratum 2**: "Qualitative correspondence with common sense." (**compatible with binary logic**)
3. **Desideratum 3**: "Consistent reasoning."
   - *Path independence:* If an answer can be calculated many ways, each should give the same answer.
   - *Non-ideological:* The reasoner does not leave out information.
   - *Equivalence:* Equivalent states of knowledge are represented by the same number.

$\Rightarrow$ Based on 3 desiderata, Jaynes mathematically proves that probability theory **and the maximum entropy principle** are the only consistent theory of plausible reasoning.

From SciFi to maths: Jaynes' probability as extended logic



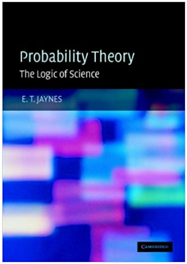
Probability Theory: the Logic of Science: In 2003, Jaynes **theorizes plausible reasoning**

1. **Desideratum 1**: "Plausibility is represented by continuous real numbers."
2. **Desideratum 2**: "Qualitative correspondence with common sense." (**compatible with binary logic**)
3. **Desideratum 3**: "Consistent reasoning."
   - *Path independence:* If an answer can be calculated many ways, each should give the same answer.
   - *Non-ideological:* The reasoner does not leave out information.
   - *Equivalence:* Equivalent states of knowledge are represented by the same number.

⇒ Based on 3 desiderata, Jaynes mathematically proves that probability theory **and the maximum entropy principle** are the only consistent theory of plausible reasoning.

In this class, we will use probability theory to "theorize fair decision making"

From SciFi to maths: Jaynes' probability as extended logic



Probability Theory: the Logic of Science: In 2003, Jaynes **theorizes plausible reasoning**

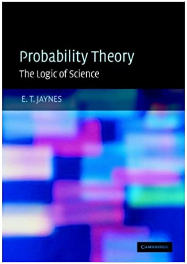
1. **Desideratum 1**: "Plausibility is represented by continuous real numbers."
2. **Desideratum 2**: "Qualitative correspondence with common sense." (**compatible with binary logic**)
3. **Desideratum 3**: "Consistent reasoning."
   - *Path independence:* If an answer can be calculated many ways, each should give the same answer.
   - *Non-ideological:* The reasoner does not leave out information.
   - *Equivalence:* Equivalent states of knowledge are represented by the same number.

$\Rightarrow$ Based on 3 desiderata, Jaynes mathematically proves that probability theory **and the maximum entropy principle** are the only consistent theory of plausible reasoning.

In this class, we will use probability theory to "theorize fair decision making"

- (again) we will also meet 3 desiderata for fairness in AI "robots"

From SciFi to maths: Jaynes' probability as extended logic



Probability Theory: the Logic of Science: In 2003, Jaynes **theorizes plausible reasoning**

1. **Desideratum 1**: "Plausibility is represented by continuous real numbers."
2. **Desideratum 2**: "Qualitative correspondence with common sense." (**compatible with binary logic**)
3. **Desideratum 3**: "Consistent reasoning."
   - *Path independence:* If an answer can be calculated many ways, each should give the same answer.
   - *Non-ideological:* The reasoner does not leave out information.
   - *Equivalence:* Equivalent states of knowledge are represented by the same number.

$\Rightarrow$ Based on 3 desiderata, Jaynes mathematically proves that probability theory **and the maximum entropy principle** are the only consistent theory of plausible reasoning.

In this class, we will use probability theory to "theorize fair decision making"

- (again) we will also meet 3 desiderata for fairness in AI "robots"
- (but again) these will fail to be satisfying as mutually incompatible (unless in trivial cases)

What's so complicated with fairness in AI?

Chicken and egg problem:

What's so complicated with fairness in AI?

Chicken and egg problem:

- a machine/algorithm is fully objective, follows a sequence of requests (Second Law of Robots: it "obeys the orders given it by human beings")

What's so complicated with fairness in AI?

Chicken and egg problem:

- a machine/algorithm is fully objective, follows a sequence of requests (Second Law of Robots: it "obeys the orders given it by human beings")
  $\Rightarrow$ They serve society better than biased humans.

What's so complicated with fairness in AI?

Chicken and egg problem:

- a machine/algorithm is fully objective, follows a sequence of requests (Second Law of Robots: it "obeys the orders given it by human beings")
  $\Rightarrow$ They serve society better than biased humans.

- **yes, but. . . ,**

What's so complicated with fairness in AI?

Chicken and egg problem:

- a machine/algorithm is fully objective, follows a sequence of requests (Second Law of Robots: it "obeys the orders given it by human beings")
  $\Rightarrow$ They serve society better than biased humans.

- **yes, but. . . ,** sequence of requests entered by subjective human beings

What's so complicated with fairness in AI?

Chicken and egg problem:

- a machine/algorithm is fully objective, follows a sequence of requests (Second Law of Robots: it "obeys the orders given it by human beings")
  $\Rightarrow$ They serve society better than biased humans.

- **yes, but. . . ,** sequence of requests entered by subjective human beings
  $\Rightarrow$ Algorithms **transfer** the biases of human beings.

What's so complicated with fairness in AI?

Chicken and egg problem:

- a machine/algorithm is fully objective, follows a sequence of requests (Second Law of Robots: it "obeys the orders given it by human beings")
  $\Rightarrow$ They serve society better than biased humans.
- **yes, but. . . ,** sequence of requests entered by subjective human beings
  $\Rightarrow$ Algorithms **transfer** the biases of human beings.
- **and. . . , it gets worse:**

What's so complicated with fairness in AI?

Chicken and egg problem:

- a machine/algorithm is fully objective, follows a sequence of requests (Second Law of Robots: it "obeys the orders given it by human beings")
  $\Rightarrow$ They serve society better than biased humans.

- **yes, but. . . ,** sequence of requests entered by subjective human beings
  $\Rightarrow$ Algorithms **transfer** the biases of human beings.

- **and. . . , it gets worse:** biased decisions in turn bias the future datasets used to refine the algorithms

What's so complicated with fairness in AI?

Chicken and egg problem:

- a machine/algorithm is fully objective, follows a sequence of requests (Second Law of Robots: it "obeys the orders given it by human beings")
  ⇒ They serve society better than biased humans.
- **yes, but. . . ,** sequence of requests entered by subjective human beings
  ⇒ Algorithms **transfer** the biases of human beings.
- **and. . . , it gets worse:** biased decisions in turn bias the future datasets used to refine the algorithms
  ⇒ Algorithms **reinforce** the biases of human beings.

What's so complicated with fairness in AI?

Chicken and egg problem:

- a machine/algorithm is fully objective, follows a sequence of requests (Second Law of Robots: it "obeys the orders given it by human beings")
  $\Rightarrow$ They serve society better than biased humans.

- **yes, but. . . ,** sequence of requests entered by subjective human beings
  $\Rightarrow$ Algorithms **transfer** the biases of human beings.

- **and. . . , it gets worse:** biased decisions in turn bias the future datasets used to refine the algorithms
  $\Rightarrow$ Algorithms **reinforce** the biases of human beings.

- **and. . . , it gets even worse:**

What's so complicated with fairness in AI?

Chicken and egg problem:

- a machine/algorithm is fully objective, follows a sequence of requests (Second Law of Robots: it "obeys the orders given it by human beings")
  $\Rightarrow$ They serve society better than biased humans.
- **yes, but. . . ,** sequence of requests entered by subjective human beings
  $\Rightarrow$ Algorithms **transfer** the biases of human beings.
- **and. . . , it gets worse:** biased decisions in turn bias the future datasets used to refine the algorithms
  $\Rightarrow$ Algorithms **reinforce** the biases of human beings.
- **and. . . , it gets even worse:** humans interventions are limited:

What's so complicated with fairness in AI?

Chicken and egg problem:

- a machine/algorithm is fully objective, follows a sequence of requests (Second Law of Robots: it "obeys the orders given it by human beings")
  $\Rightarrow$ They serve society better than biased humans.

- **yes, but. . . ,** sequence of requests entered by subjective human beings
  $\Rightarrow$ Algorithms **transfer** the biases of human beings.

- **and. . . , it gets worse:** biased decisions in turn bias the future datasets used to refine the algorithms
  $\Rightarrow$ Algorithms **reinforce** the biases of human beings.

- **and. . . , it gets even worse:** humans interventions are limited:
  - we trust the objectivity of algorithms (they obey, and cannot go wrong)

What's so complicated with fairness in AI?

Chicken and egg problem:

- a machine/algorithm is fully objective, follows a sequence of requests (Second Law of Robots: it "obeys the orders given it by human beings")
  $\Rightarrow$ They serve society better than biased humans.

- **yes, but. . . ,** sequence of requests entered by subjective human beings
  $\Rightarrow$ Algorithms **transfer** the biases of human beings.

- **and. . . , it gets worse:** biased decisions in turn bias the future datasets used to refine the algorithms
  $\Rightarrow$ Algorithms **reinforce** the biases of human beings.

- **and. . . , it gets even worse:** humans interventions are limited:
  - we trust the objectivity of algorithms (they obey, and cannot go wrong)
  - algorithms now are black boxes: we do not know how they treat the data

What's so complicated with fairness in AI?

Chicken and egg problem:

- a machine/algorithm is fully objective, follows a sequence of requests (Second Law of Robots: it "obeys the orders given it by human beings")
  $\Rightarrow$ They serve society better than biased humans.
- **yes, but...,** sequence of requests entered by subjective human beings
  $\Rightarrow$ Algorithms **transfer** the biases of human beings.
- **and..., it gets worse:** biased decisions in turn bias the future datasets used to refine the algorithms
  $\Rightarrow$ Algorithms **reinforce** the biases of human beings.
- **and..., it gets even worse:** humans interventions are limited:
  - we trust the objectivity of algorithms (they obey, and cannot go wrong)
  - algorithms now are black boxes: we do not know how they treat the data

Consequence: open door to unfair decisions, uncontrollable behavior, unseen biases.

Illustration: search engines “human-friendly” behavior

Search engines: AI-improved to help people easily find their search

Illustration: search engines “human-friendly” behavior

Search engines: AI-improved to help people easily find their search

- (level 0) remembers previous searches

Illustration: search engines “human-friendly” behavior

Search engines: AI-improved to help people easily find their search

- (level 0) remembers previous searches
  $\Rightarrow$ creates focus on already known information

Illustration: search engines “human-friendly” behavior

Search engines: AI-improved to help people easily find their search

- (level 0) remembers previous searches
  $\Rightarrow$ creates focus on already known information
  $\Rightarrow$ less priority on opposite opinions, other information

Illustration: search engines “human-friendly” behavior

Search engines: AI-improved to help people easily find their search

- (level 0) remembers previous searches
  $\Rightarrow$ creates focus on already known information
  $\Rightarrow$ less priority on opposite opinions, other information

- (next level) infers preferences (tries to anticipate your search)

Illustration: search engines “human-friendly” behavior

Search engines: AI-improved to help people easily find their search

- (level 0) remembers previous searches
  ⇒ creates focus on already known information
  ⇒ less priority on opposite opinions, other information

- (next level) infers preferences (tries to anticipate your search)
  ⇒ in a way, algorithms **dictate** our behavior

Illustration: search engines "human-friendly" behavior

Search engines: AI-improved to help people easily find their search

- (level 0) remembers previous searches
  ⇒ creates focus on already known information
  ⇒ less priority on opposite opinions, other information

- (next level) infers preferences (tries to anticipate your search)
  ⇒ in a way, algorithms **dictate** our behavior
  ⇒ being often black-boxes, difficult to know what this really does

Illustration: search engines “human-friendly” behavior

Search engines: AI-improved to help people easily find their search

- (level 0) remembers previous searches
  $\Rightarrow$ creates focus on already known information
  $\Rightarrow$ less priority on opposite opinions, other information

- (next level) infers preferences (tries to anticipate your search)
  $\Rightarrow$ in a way, algorithms **dictate** our behavior
  $\Rightarrow$ being often black-boxes, difficult to know what this really does
  $\Rightarrow$ algorithm often based on “best effort” **(represents majority)**: homogeneous behavior enforced!

Illustration: search engines “human-friendly” behavior

Search engines: AI-improved to help people easily find their search

- (level 0) remembers previous searches
  $\Rightarrow$ creates focus on already known information
  $\Rightarrow$ less priority on opposite opinions, other information
- (next level) infers preferences (tries to anticipate your search)
  $\Rightarrow$ in a way, algorithms **dictate** our behavior
  $\Rightarrow$ being often black-boxes, difficult to know what this really does
  $\Rightarrow$ algorithm often based on “best effort” **(represents majority)**: homogeneous behavior enforced!

Ethical, law issues:

Illustration: search engines "human-friendly" behavior

Search engines: AI-improved to help people easily find their search

- (level 0) remembers previous searches
  ⇒ creates focus on already known information
  ⇒ less priority on opposite opinions, other information
- (next level) infers preferences (tries to anticipate your search)
  ⇒ in a way, algorithms **dictate** our behavior
  ⇒ being often black-boxes, difficult to know what this really does
  ⇒ algorithm often based on "best effort" **(represents majority)**: homogeneous behavior enforced!

Ethical, law issues:

- polarization of information (reinforcement of majority choices)

Illustration: search engines “human-friendly” behavior

Search engines: AI-improved to help people easily find their search

- (level 0) remembers previous searches
  ⇒ creates focus on already known information
  ⇒ less priority on opposite opinions, other information

- (next level) infers preferences (tries to anticipate your search)
  ⇒ in a way, algorithms **dictate** our behavior
  ⇒ being often black-boxes, difficult to know what this really does
  ⇒ algorithm often based on “best effort” **(represents majority)**: homogeneous behavior enforced!

Ethical, law issues:

- polarization of information (reinforcement of majority choices)
- biases can be introduced **in** the machine, or **by** the machine

Illustration: search engines “human-friendly” behavior

Search engines: AI-improved to help people easily find their search

- (level 0) remembers previous searches
  $\Rightarrow$ creates focus on already known information
  $\Rightarrow$ less priority on opposite opinions, other information

- (next level) infers preferences (tries to anticipate your search)
  $\Rightarrow$ in a way, algorithms **dictate** our behavior
  $\Rightarrow$ being often black-boxes, difficult to know what this really does
  $\Rightarrow$ algorithm often based on “best effort” **(represents majority)**: homogeneous behavior enforced!

Ethical, law issues:

- polarization of information (reinforcement of majority choices)
- biases can be introduced **in** the machine, or **by** the machine
- inequity of information access in minority populations.

Illustration: gender discrimination

Typical scenario: Automated recruitment, ad proposal, etc.,

Illustration: gender discrimination

Typical scenario: Automated recruitment, ad proposal, etc.,

1. Machine learnt to identify men from women (selective target of the job/ad) from database

Illustration: gender discrimination

Typical scenario: Automated recruitment, ad proposal, etc.,

1. Machine learnt to identify men from women (selective target of the job/ad) from database

2. Human operator does not know how the machine proceeds

Illustration: gender discrimination

Typical scenario: Automated recruitment, ad proposal, etc.,

1. Machine learnt to identify men from women (selective target of the job/ad) from database

2. Human operator does not know how the machine proceeds
   In reality, machine uses features: hair-length, lip color, presence of earrings, etc.

Illustration: gender discrimination

Typical scenario: Automated recruitment, ad proposal, etc.,

1. Machine learnt to identify men from women (selective target of the job/ad) from database

2. Human operator does not know how the machine proceeds
   In reality, machine uses features: hair-length, lip color, presence of earrings, etc.

3. Following people ill-classified: do not receive ad, job proposal

Illustration: gender discrimination

Typical scenario: Automated recruitment, ad proposal, etc.,

1. Machine learnt to identify men from women (selective target of the job/ad) from database

2. Human operator does not know how the machine proceeds
   In reality, machine uses features: hair-length, lip color, presence of earrings, etc.
3. Following people ill-classified: do not receive ad, job proposal

Consequence: clear example of undesired/uncontrolled discrimination:

Illustration: gender discrimination

Typical scenario: Automated recruitment, ad proposal, etc.,

1. Machine learnt to identify men from women (selective target of the job/ad) from database

2. Human operator does not know how the machine proceeds
   In reality, machine uses features: hair-length, lip color, presence of earrings, etc.
3. Following people ill-classified: do not receive ad, job proposal

Consequence: clear example of undesired/uncontrolled discrimination:

- unfairness to several minorities

Illustration: gender discrimination

Typical scenario: Automated recruitment, ad proposal, etc.,

1. Machine learnt to identify men from women (selective target of the job/ad) from database

2. Human operator does not know how the machine proceeds
   In reality, machine uses features: hair-length, lip color, presence of earrings, etc.
3. Following people ill-classified: do not receive ad, job proposal

Consequence: clear example of undesired/uncontrolled discrimination:

- unfairness to several minorities
- hard to anticipate (even with larger database, minorities won't alter features!)

Illustration: gender discrimination

Typical scenario: Automated recruitment, ad proposal, etc.,

1. Machine learnt to identify men from women (selective target of the job/ad) from database

2. Human operator does not know how the machine proceeds
   In reality, machine uses features: hair-length, lip color, presence of earrings, etc.
3. Following people ill-classified: do not receive ad, job proposal

Consequence: clear example of undesired/uncontrolled discrimination:

- unfairness to several minorities
- hard to anticipate (even with larger database, minorities won't alter features!)
- hard to defend on basis of law

Illustration: gender discrimination

Under SVM formulation: in best effort strategy, minority groups excluded from optimization

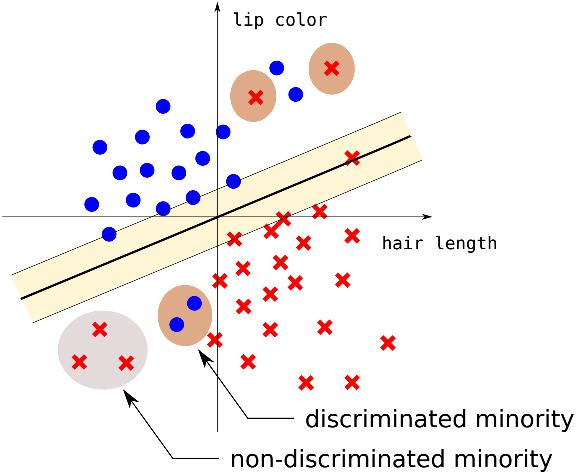
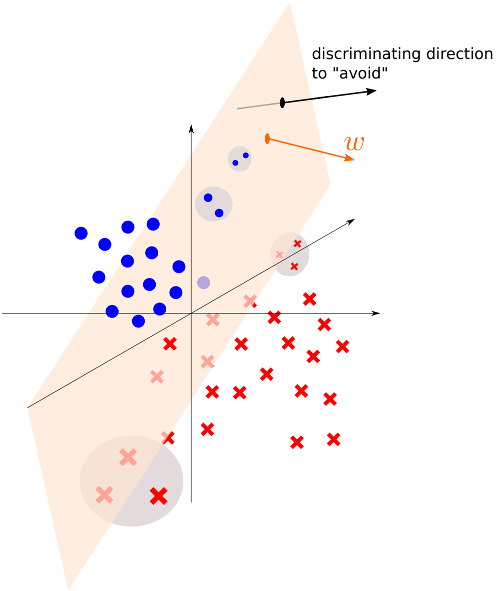


Illustration: gender discrimination

Under SVM formulation: possible counter-measure: force separating hyperplane against discriminating directions?



Main objectives and messages of the class

Main objectives:

Main objectives and messages of the class

Main objectives:

1. connect machine learning concepts to law and ethics

Main objectives and messages of the class

Main objectives:

1. connect machine learning concepts to law and ethics
2. identify **desiderata of ethical, fair machine learning**

Main objectives and messages of the class

Main objectives:

1. connect machine learning concepts to law and ethics
2. identify **desiderata of ethical, fair machine learning**
3. **mathematically formalize** the notion of fairness

Main objectives and messages of the class

Main objectives:

1. connect machine learning concepts to law and ethics
2. identify **desiderata of ethical, fair machine learning**
3. **mathematically formalize** the notion of fairness

Take-home messages:

Main objectives and messages of the class

Main objectives:

1. connect machine learning concepts to law and ethics
2. identify **desiderata of ethical, fair machine learning**
3. **mathematically formalize** the notion of fairness

Take-home messages:

1. fairness in AI is a **nascent field**: still on shaky grounds!

Main objectives and messages of the class

Main objectives:

1. connect machine learning concepts to law and ethics
2. identify **desiderata of ethical, fair machine learning**
3. **mathematically formalize** the notion of fairness

Take-home messages:

1. fairness in AI is a **nascent field**: still on shaky grounds!
2. recent mathematical formalization on basic proba/information theory grounds

Main objectives and messages of the class

Main objectives:

1. connect machine learning concepts to law and ethics
2. identify **desiderata of ethical, fair machine learning**
3. **mathematically formalize** the notion of fairness

Take-home messages:

1. fairness in AI is a **nascent field**: still on shaky grounds!
2. recent mathematical formalization on basic proba/information theory grounds
3. we will exhibit three “laws of fairness AI” under the form of desiderata

Main objectives and messages of the class

Main objectives:

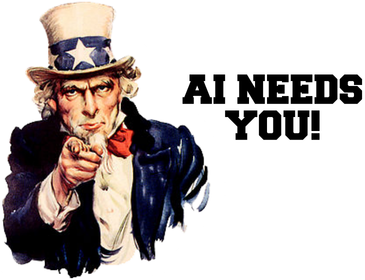
1. connect machine learning concepts to law and ethics
2. identify **desiderata of ethical, fair machine learning**
3. **mathematically formalize** the notion of fairness

Take-home messages:

1. fairness in AI is a **nascent field**: still on shaky grounds!
2. recent mathematical formalization on basic proba/information theory grounds
3. we will exhibit three "laws of fairness AI" under the form of desiderata
4. **Big problem**: three desiderata mutually incompatible!

Main objectives and messages of the class

. . . incomplete conclusion . . . : as future AI engineers, you will be the ambassadors of **a fair AI**



About the material

This class: strongly inspired by works of Solon Barocas, Moritz Hardt, Arvind Narayanan.

About the material

This class: strongly inspired by works of Solon Barocas, Moritz Hardt, Arvind Narayanan.

Recommended lectures/videos:

About the material

This class: strongly inspired by works of Solon Barocas, Moritz Hardt, Arvind Narayanan.

Recommended lectures/videos:

- NeurIPS 2017 – Tutorial (video): https://fairmlbook.org/tutorial1.html

About the material

This class: strongly inspired by works of Solon Barocas, Moritz Hardt, Arvind Narayanan.

Recommended lectures/videos:

- NeurIPS 2017 – Tutorial (video): https://fairmlbook.org/tutorial1.html
- “Fairness and machine learning” online book: https://fairmlbook.org/

About the material

This class: strongly inspired by works of Solon Barocas, Moritz Hardt, Arvind Narayanan.

Recommended lectures/videos:

- NeurIPS 2017 – Tutorial (video): https://fairmlbook.org/tutorial1.html
- "Fairness and machine learning" online book: https://fairmlbook.org/
- related material (just google-scholar "fairness machine learning")

Outline

Fairness: law and ethics

The question of discriminating data and information

The problem: algorithms reinforce human prejudices

The question of discriminating data and information

The problem: algorithms reinforce human prejudices

- algorithms written and maintained by people,

The question of discriminating data and information

The problem: algorithms reinforce human prejudices

- algorithms written and maintained by people,
- data labelled and selected (even passively) by people.

The question of discriminating data and information

The problem: algorithms reinforce human prejudices

- algorithms written and maintained by people,
- data labelled and selected (even passively) by people.
  (passive selection: minorities in groups not answering polls)

The question of discriminating data and information

The problem: algorithms reinforce human prejudices

- algorithms written and maintained by people,
- data labelled and selected (even passively) by people.
  (passive selection: minorities in groups not answering polls)
- snowballing effect if data appended by outputs of previous algorithms

The question of discriminating data and information

The problem: algorithms reinforce human prejudices

- algorithms written and maintained by people,
- data labelled and selected (even passively) by people.
  (passive selection: minorities in groups not answering polls)
- snowballing effect if data appended by outputs of previous algorithms

Solution: actively account for biases

The question of discriminating data and information

The problem: algorithms reinforce human prejudices

- algorithms written and maintained by people,
- data labelled and selected (even passively) by people.
  (passive selection: minorities in groups not answering polls)
- snowballing effect if data appended by outputs of previous algorithms

Solution: actively account for biases

- in itself an ethical problem:
  - admit existence of minorities
  - treat minorities differently

The question of discriminating data and information

The problem: algorithms reinforce human prejudices

- algorithms written and maintained by people,
- data labelled and selected (even passively) by people.
  (passive selection: minorities in groups not answering polls)
- snowballing effect if data appended by outputs of previous algorithms

Solution: actively account for biases

- in itself an ethical problem:
  - admit existence of minorities
  - treat minorities differently
- paradox of the use of discriminative information:

The question of discriminating data and information

The problem: algorithms reinforce human prejudices

- algorithms written and maintained by people,
- data labelled and selected (even passively) by people.
  (passive selection: minorities in groups not answering polls)
- snowballing effect if data appended by outputs of previous algorithms

Solution: actively account for biases

- in itself an ethical problem:
  - admit existence of minorities
  - treat minorities differently
- paradox of the use of discriminative information:
  - exploiting private information helps avoid discrimination

The question of discriminating data and information

The problem: algorithms reinforce human prejudices

- algorithms written and maintained by people,
- data labelled and selected (even passively) by people.
  (passive selection: minorities in groups not answering polls)
- snowballing effect if data appended by outputs of previous algorithms

Solution: actively account for biases

- in itself an ethical problem:
  - admit existence of minorities
  - treat minorities differently
- paradox of the use of discriminative information:
  - exploiting private information helps avoid discrimination
    $\Rightarrow$ Discriminative sensitive information needed!

The question of discriminating data and information

The problem: algorithms reinforce human prejudices

- algorithms written and maintained by people,
- data labelled and selected (even passively) by people.
  (passive selection: minorities in groups not answering polls)
- snowballing effect if data appended by outputs of previous algorithms

Solution: actively account for biases

- in itself an ethical problem:
  - admit existence of minorities
  - treat minorities differently
- paradox of the use of discriminative information:
  - exploiting private information helps avoid discrimination
    $\Rightarrow$ Discriminative sensitive information needed!
  - retrieve private information is unethical
    (possibility of bad intentional usage)

The question of discriminating data and information

The problem: algorithms reinforce human prejudices

- algorithms written and maintained by people,
- data labelled and selected (even passively) by people.
  (passive selection: minorities in groups not answering polls)
- snowballing effect if data appended by outputs of previous algorithms

Solution: actively account for biases

- in itself an ethical problem:
  - admit existence of minorities
  - treat minorities differently
- paradox of the use of discriminative information:
  - exploiting private information helps avoid discrimination
    $\Rightarrow$ Discriminative sensitive information needed!
  - retrieve private information is unethical
    (possibility of bad intentional usage)
  - indirect sensitive information inference is also unethical...

Disparate treatment and disparate impact

Two legal difficulties:

discriminating data $\Leftrightarrow$ **disparate treatment**

Disparate treatment and disparate impact

Two legal difficulties:

discriminating data $\Leftrightarrow$ **disparate treatment**

discriminating algorithms $\Leftrightarrow$ **disparate impact**

Disparate treatment and disparate impact

Two legal difficulties:

discriminating data $\Leftrightarrow$ **disparate treatment**

discriminating algorithms $\Leftrightarrow$ **disparate impact**

Disparate treatment:

Disparate treatment and disparate impact

Two legal difficulties:

discriminating data $\Leftrightarrow$ **disparate treatment**

discriminating algorithms $\Leftrightarrow$ **disparate impact**

Disparate treatment:

- laws exist that protect subgroups against discrimination

Disparate treatment and disparate impact

Two legal difficulties:

discriminating data $\Leftrightarrow$ **disparate treatment**

discriminating algorithms $\Leftrightarrow$ **disparate impact**

Disparate treatment:

- laws exist that protect subgroups against discrimination
- the very fact of *using discriminating information* is illegal

Disparate treatment and disparate impact

Two legal difficulties:

discriminating data $\Leftrightarrow$ **disparate treatment**

discriminating algorithms $\Leftrightarrow$ **disparate impact**

Disparate treatment:

- laws exist that protect subgroups against discrimination
- the very fact of *using discriminating information* is illegal
- even if it has no impact!

Disparate treatment and disparate impact

Two legal difficulties:

discriminating data	$\Leftrightarrow$	**disparate treatment**
discriminating algorithms	$\Leftrightarrow$	**disparate impact**

Disparate treatment:

- laws exist that protect subgroups against discrimination
- the very fact of *using discriminating information* is illegal
- even if it has no impact!

- exploiting **proxies** to target these classes intentionally is also illegal

Disparate treatment and disparate impact

Two legal difficulties:

discriminating data $\Leftrightarrow$ **disparate treatment**

discriminating algorithms $\Leftrightarrow$ **disparate impact**

Disparate treatment:

- laws exist that protect subgroups against discrimination
- the very fact of *using discriminating information* is illegal
- even if it has no impact!

- exploiting **proxies** to target these classes intentionally is also illegal
  (e.g., name, zip codes, places of residence to identify minorities)

Disparate treatment and disparate impact

Two legal difficulties:

discriminating data $\Leftrightarrow$ **disparate treatment**

discriminating algorithms $\Leftrightarrow$ **disparate impact**

Disparate treatment:

- laws exist that protect subgroups against discrimination
- the very fact of *using discriminating information* is illegal
- even if it has no impact!

- exploiting **proxies** to target these classes intentionally is also illegal
  (e.g., name, zip codes, places of residence to identify minorities)

Disparate impact:

Disparate treatment and disparate impact

Two legal difficulties:

discriminating data	$\Leftrightarrow$	**disparate treatment**
discriminating algorithms	$\Leftrightarrow$	**disparate impact**

Disparate treatment:

- laws exist that protect subgroups against discrimination
- the very fact of *using discriminating information* is illegal
- even if it has no impact!

- exploiting **proxies** to target these classes intentionally is also illegal
  (e.g., name, zip codes, places of residence to identify minorities)

Disparate impact:

- consists in using features not intentionally favoring a class

Disparate treatment and disparate impact

Two legal difficulties:

discriminating data	$\Leftrightarrow$	**disparate treatment**
discriminating algorithms	$\Leftrightarrow$	**disparate impact**

Disparate treatment:

- laws exist that protect subgroups against discrimination
- the very fact of *using discriminating information* is illegal
- even if it has no impact!

- exploiting **proxies** to target these classes intentionally is also illegal
  (e.g., name, zip codes, places of residence to identify minorities)

Disparate impact:

- consists in using features not intentionally favoring a class
- this is legal so long that the process used to reach the outcome **is justified**

Disparate treatment and disparate impact

Two legal difficulties:

discriminating data $\Leftrightarrow$ **disparate treatment**

discriminating algorithms $\Leftrightarrow$ **disparate impact**

Disparate treatment:

- laws exist that protect subgroups against discrimination
- the very fact of *using discriminating information* is illegal
- even if it has no impact!
- exploiting **proxies** to target these classes intentionally is also illegal
  (e.g., name, zip codes, places of residence to identify minorities)

Disparate impact:

- consists in using features not intentionally favoring a class
- this is legal so long that the process used to reach the outcome **is justified**
- question to be asked: is it avoidable?

Disparate impact in law

The law in the US: typical lawsuit process

Disparate impact in law

The law in the US: typical lawsuit process

1. plaintiff of discrimination (say job recruitment automated process) needs to prove 20% disparity between minority/majority groups

Disparate impact in law

The law in the US: typical lawsuit process

1. plaintiff of discrimination (say job recruitment automated process) needs to prove 20% disparity between minority/majority groups
2. defendant must prove the method is necessary (unavoidable) to reach sought target (e.g., specificities of a job)

Disparate impact in law

The law in the US: typical lawsuit process

1. plaintiff of discrimination (say job recruitment automated process) needs to prove 20% disparity between minority/majority groups
2. defendant must prove the method is necessary (unavoidable) to reach sought target (e.g., specificities of a job)
3. plaintiff must then provide less ($< 20\%$) discriminative alternative.

Disparate impact in law

The law in the US: typical lawsuit process

1. plaintiff of discrimination (say job recruitment automated process) needs to prove 20% disparity between minority/majority groups
2. defendant must prove the method is necessary (unavoidable) to reach sought target (e.g., specificities of a job)
3. plaintiff must then provide less ($< 20\%$) discriminative alternative.

Example: job application on construction site

Disparate impact in law

The law in the US: typical lawsuit process

1. plaintiff of discrimination (say job recruitment automated process) needs to prove 20% disparity between minority/majority groups
2. defendant must prove the method is necessary (unavoidable) to reach sought target (e.g., specificities of a job)
3. plaintiff must then provide less ($< 20\%$) discriminative alternative.

Example: job application on construction site

1. plaintiff complaint: job questionnaire asked for "maximum load heaved by applicant", which favors men more than 20%

Disparate impact in law

The law in the US: typical lawsuit process

1. plaintiff of discrimination (say job recruitment automated process) needs to prove 20% disparity between minority/majority groups
2. defendant must prove the method is necessary (unavoidable) to reach sought target (e.g., specificities of a job)
3. plaintiff must then provide less ($< 20\%$) discriminative alternative.

Example: job application on construction site

1. plaintiff complaint: job questionnaire asked for "maximum load heaved by applicant", which favors men more than 20%
2. defendant claim: necessary question to assess employee ability to the job

Disparate impact in law

The law in the US: typical lawsuit process

1. plaintiff of discrimination (say job recruitment automated process) needs to prove 20% disparity between minority/majority groups
2. defendant must prove the method is necessary (unavoidable) to reach sought target (e.g., specificities of a job)
3. plaintiff must then provide less ($< 20\%$) discriminative alternative.

Example: job application on construction site

1. plaintiff complaint: job questionnaire asked for "maximum load heaved by applicant", which favors men more than 20%
2. defendant claim: necessary question to assess employee ability to the job
3. plaintiff may retort: live tests with modern construction site equipment has same effect, but is less discriminating.

Fighting disparate treatment and impact

Fighting disparate treatment:

Fighting disparate treatment and impact

Fighting disparate treatment:

- impose **procedural fairness** (only exploit data about worthiness, directly linked to objective)

Fighting disparate treatment and impact

Fighting disparate treatment:

- impose **procedural fairness** (only exploit data about worthiness, directly linked to objective)
- target **equality of opportunity** (all individuals, or items in groups, have equal success rate, irrespective of discriminating subgroups)

Fighting disparate treatment and impact

Fighting disparate treatment:

- impose **procedural fairness** (only exploit data about worthiness, directly linked to objective)
- target **equality of opportunity** (all individuals, or items in groups, have equal success rate, irrespective of discriminating subgroups)

Fighting disparate impact:

Fighting disparate treatment and impact

Fighting disparate treatment:

- impose **procedural fairness** (only exploit data about worthiness, directly linked to objective)
- target **equality of opportunity** (all individuals, or items in groups, have equal success rate, irrespective of discriminating subgroups)

Fighting disparate impact:

- minimize inequality of treatment within subgroups

Fighting disparate treatment and impact

Fighting disparate treatment:

- impose **procedural fairness** (only exploit data about worthiness, directly linked to objective)
- target **equality of opportunity** (all individuals, or items in groups, have equal success rate, irrespective of discriminating subgroups)

Fighting disparate impact:

- minimize inequality of treatment within subgroups
- homogenize distribution of outcomes

Fighting disparate treatment and impact

Fighting disparate treatment:

- impose **procedural fairness** (only exploit data about worthiness, directly linked to objective)
- target **equality of opportunity** (all individuals, or items in groups, have equal success rate, irrespective of discriminating subgroups)

Fighting disparate impact:

- minimize inequality of treatment within subgroups
- homogenize distribution of outcomes

Overall goal: organize society such that **people of equal talents can achieve equal outcomes**

Fighting disparate treatment and impact

Fighting disparate treatment:

- impose **procedural fairness** (only exploit data about worthiness, directly linked to objective)
- target **equality of opportunity** (all individuals, or items in groups, have equal success rate, irrespective of discriminating subgroups)

Fighting disparate impact:

- minimize inequality of treatment within subgroups
- homogenize distribution of outcomes

Overall goal: organize society such that **people of equal talents can achieve equal outcomes**

Difficulty:

Fighting disparate treatment and impact

Fighting disparate treatment:

- impose **procedural fairness** (only exploit data about worthiness, directly linked to objective)
- target **equality of opportunity** (all individuals, or items in groups, have equal success rate, irrespective of discriminating subgroups)

Fighting disparate impact:

- minimize inequality of treatment within subgroups
- homogenize distribution of outcomes

Overall goal: organize society such that **people of equal talents can achieve equal outcomes**

Difficulty:

- should one account for past injustice suffered by minorities? (i.e., payback for past unequal outcomes to achieve equal "integrated outcomes"?)

Fighting disparate treatment and impact

Fighting disparate treatment:

- impose **procedural fairness** (only exploit data about worthiness, directly linked to objective)
- target **equality of opportunity** (all individuals, or items in groups, have equal success rate, irrespective of discriminating subgroups)

Fighting disparate impact:

- minimize inequality of treatment within subgroups
- homogenize distribution of outcomes

Overall goal: organize society such that **people of equal talents can achieve equal outcomes**

Difficulty:

- should one account for past injustice suffered by minorities? (i.e., payback for past unequal outcomes to achieve equal "integrated outcomes"?)
- contradicts homogeneous outcomes!

Fighting disparate treatment and impact

Fighting disparate treatment:

- impose **procedural fairness** (only exploit data about worthiness, directly linked to objective)
- target **equality of opportunity** (all individuals, or items in groups, have equal success rate, irrespective of discriminating subgroups)

Fighting disparate impact:

- minimize inequality of treatment within subgroups
- homogenize distribution of outcomes

Overall goal: organize society such that **people of equal talents can achieve equal outcomes**

Difficulty:

- should one account for past injustice suffered by minorities? (i.e., payback for past unequal outcomes to achieve equal "integrated outcomes"?)
- contradicts homogeneous outcomes!
- and to minimize disparate outcomes, one may need to know the subgroups, treat individuals differently

Fighting disparate treatment and impact

Fighting disparate treatment:

- impose **procedural fairness** (only exploit data about worthiness, directly linked to objective)
- target **equality of opportunity** (all individuals, or items in groups, have equal success rate, irrespective of discriminating subgroups)

Fighting disparate impact:

- minimize inequality of treatment within subgroups
- homogenize distribution of outcomes

Overall goal: organize society such that **people of equal talents can achieve equal outcomes**

Difficulty:

- should one account for past injustice suffered by minorities? (i.e., payback for past unequal outcomes to achieve equal "integrated outcomes"?)
- contradicts homogeneous outcomes!
- and to minimize disparate outcomes, one may need to know the subgroups, treat individuals differently
- but this contradicts disparate treatment!. . .

Fighting disparate treatment and impact

Fighting disparate treatment:

- impose **procedural fairness** (only exploit data about worthiness, directly linked to objective)
- target **equality of opportunity** (all individuals, or items in groups, have equal success rate, irrespective of discriminating subgroups)

Fighting disparate impact:

- minimize inequality of treatment within subgroups
- homogenize distribution of outcomes

Overall goal: organize society such that **people of equal talents can achieve equal outcomes**

Difficulty:

- should one account for past injustice suffered by minorities? (i.e., payback for past unequal outcomes to achieve equal “integrated outcomes”?)
- contradicts homogeneous outcomes!
- and to minimize disparate outcomes, one may need to know the subgroups, treat individuals differently
- but this contradicts disparate treatment!. . .

Consequence: Tension between disparate treatment and disparate outcomes!

Illustrating tension between disparate treatment and disparate outcomes

Job employment process:

Illustrating tension between disparate treatment and disparate outcomes

Job employment process:

1. plaintiff complaint: in job chances, procedure indirectly favors white people (e.g., university reputation, ease to reach job location, family constraint, etc.)

Illustrating tension between disparate treatment and disparate outcomes

Job employment process:

1. plaintiff complaint: in job chances, procedure indirectly favors white people (e.g., university reputation, ease to reach job location, family constraint, etc.)
   $\Rightarrow$ disparate outcome (no information on color is used)

Illustrating tension between disparate treatment and disparate outcomes

Job employment process:

1. plaintiff complaint: in job chances, procedure indirectly favors white people (e.g., university reputation, ease to reach job location, family constraint, etc.)
   $\Rightarrow$ disparate outcome (no information on color is used)

2. HR change the rule to account for "typical black people difficulties"

Illustrating tension between disparate treatment and disparate outcomes

Job employment process:

1. plaintiff complaint: in job chances, procedure indirectly favors white people (e.g., university reputation, ease to reach job location, family constraint, etc.)
   $\Rightarrow$ disparate outcome (no information on color is used)

2. HR change the rule to account for "typical black people difficulties"
   $\Rightarrow$ induces disparate treatment! (voluntary usage of color people-targeting features)

Illustrating tension between disparate treatment and disparate outcomes

Job employment process:

1. plaintiff complaint: in job chances, procedure indirectly favors white people (e.g., university reputation, ease to reach job location, family constraint, etc.)
   $\Rightarrow$ disparate outcome (no information on color is used)

2. HR change the rule to account for “typical black people difficulties”
   $\Rightarrow$ induces disparate treatment! (voluntary usage of color people-targeting features)

3. white people in turn complain: job chances have become unequal!

Outline

How machines learn to discriminate

How machines learn to discriminate

Skewed samples: a vicious cycle!

How machines learn to discriminate

Skewed samples: a vicious cycle!

1. unfair machines **bias the decision maker** (the human)

How machines learn to discriminate

Skewed samples: a vicious cycle!

1. unfair machines **bias the decision maker** (the human)
2. future observations (made by the biased decision maker) will **confirm the bias**

How machines learn to discriminate

Skewed samples: a vicious cycle!

1. unfair machines **bias the decision maker** (the human)
2. future observations (made by the biased decision maker) will **confirm the bias**
3. this **reduces opportunities to see instances contradicting the bias**

How machines learn to discriminate

Skewed samples: a vicious cycle!

1. unfair machines **bias the decision maker** (the human)
2. future observations (made by the biased decision maker) will **confirm the bias**
3. this **reduces opportunities to see instances contradicting the bias**

Skewed samples: the example of crimes:

How machines learn to discriminate

Skewed samples: a vicious cycle!

1. unfair machines **bias the decision maker** (the human)
2. future observations (made by the biased decision maker) will **confirm the bias**
3. this **reduces opportunities to see instances contradicting the bias**

Skewed samples: the example of crimes:

1. a machine says that black people are more likely to commit crimes, making decision maker (the police) take action on blacks

How machines learn to discriminate

Skewed samples: a vicious cycle!

1. unfair machines **bias the decision maker** (the human)
2. future observations (made by the biased decision maker) will **confirm the bias**
3. this **reduces opportunities to see instances contradicting the bias**

Skewed samples: the example of crimes:

1. a machine says that black people are more likely to commit crimes, making decision maker (the police) take action on blacks
2. the police arrest **more** black people and less white people, reinforcing the bias

How machines learn to discriminate

Skewed samples: a vicious cycle!

1. unfair machines **bias the decision maker** (the human)
2. future observations (made by the biased decision maker) will **confirm the bias**
3. this **reduces opportunities to see instances contradicting the bias**

Skewed samples: the example of crimes:

1. a machine says that black people are more likely to commit crimes, making decision maker (the police) take action on blacks
2. the police arrest **more** black people and less white people, reinforcing the bias
3. the data feed the machine for further evaluation and decision-making, creating a **vicious cycle**.

How machines learn to discriminate

Tainted samples: i.e., bad labels

How machines learn to discriminate

Tainted samples: i.e., bad labels

- can be due to prediction based on past human decisions (or machine decisions made by humans)

Tainted samples: i.e., bad labels

- can be due to prediction based on past human decisions (or machine decisions made by humans)
- how to avoid this? $\Rightarrow$ Change the decision making.

How machines learn to discriminate

Tainted samples: i.e., bad labels

- can be due to prediction based on past human decisions (or machine decisions made by humans)
- how to avoid this? $\Rightarrow$ Change the decision making.

Tainted samples: the example of job recruitment

How machines learn to discriminate

Tainted samples: i.e., bad labels

- can be due to prediction based on past human decisions (or machine decisions made by humans)
- how to avoid this? $\Rightarrow$ Change the decision making.

Tainted samples: the example of job recruitment

- labels affected to minority subgroups by humans: were people hired?

How machines learn to discriminate

Tainted samples: i.e., bad labels

- can be due to prediction based on past human decisions (or machine decisions made by humans)
- how to avoid this? $\Rightarrow$ Change the decision making.

Tainted samples: the example of job recruitment

- labels affected to minority subgroups by humans: were people hired?
- change of decision making: how did they do in previous jobs?

How machines learn to discriminate

Tainted samples: i.e., bad labels

- can be due to prediction based on past human decisions (or machine decisions made by humans)
- how to avoid this? $\Rightarrow$ Change the decision making.

Tainted samples: the example of job recruitment

- labels affected to minority subgroups by humans: were people hired?
- change of decision making: how did they do in previous jobs?
- but still limited: exploits previous managers' biases

How machines learn to discriminate

Limited features: features less informative or less reliably collected on parts of the population

How machines learn to discriminate

Limited features: features less informative or less reliably collected on parts of the population

- typical case: good predictions for majority, weak predictions for minority

How machines learn to discriminate

Limited features: features less informative or less reliably collected on parts of the population

- typical case: good predictions for majority, weak predictions for minority (different additional problem to number of samples)

How machines learn to discriminate

Limited features: features less informative or less reliably collected on parts of the population

- typical case: good predictions for majority, weak predictions for minority (different additional problem to number of samples)
- consequence: uneven distribution of errors across population, even in equal number of samples.

How machines learn to discriminate

Limited features: features less informative or less reliably collected on parts of the population

- typical case: good predictions for majority, weak predictions for minority (different additional problem to number of samples)
- consequence: uneven distribution of errors across population, even in equal number of samples.

Limited features: data collection across wealthy vs. poor communities

How machines learn to discriminate

Limited features: features less informative or less reliably collected on parts of the population

- typical case: good predictions for majority, weak predictions for minority (different additional problem to number of samples)
- consequence: uneven distribution of errors across population, even in equal number of samples.

Limited features: data collection across wealthy vs. poor communities

- data collection medium: Internet access, access opportunity, time availability to data collection

How machines learn to discriminate

Limited features: features less informative or less reliably collected on parts of the population

- typical case: good predictions for majority, weak predictions for minority (different additional problem to number of samples)
- consequence: uneven distribution of errors across population, even in equal number of samples.

Limited features: data collection across wealthy vs. poor communities

- data collection medium: Internet access, access opportunity, time availability to data collection
- quality of information: average education level to answer polls, absence of answers when inappropriate

How machines learn to discriminate

Sample size disparity: high samples implies lower error, higher confidence

How machines learn to discriminate

Sample size disparity: high samples implies lower error, higher confidence

- **small groups have higher variance**, higher error levels

How machines learn to discriminate

Sample size disparity: high samples implies lower error, higher confidence

- **small groups have higher variance**, higher error levels
- in best effort mechanisms, smaller groups ignored to favor majority score

How machines learn to discriminate

Sample size disparity: high samples implies lower error, higher confidence

- **small groups have higher variance**, higher error levels
- in best effort mechanisms, smaller groups ignored to favor majority score

Proxies: features naturally correlated with class membership (bias in features)

How machines learn to discriminate

Sample size disparity: high samples implies lower error, higher confidence

- **small groups have higher variance**, higher error levels
- in best effort mechanisms, smaller groups ignored to favor majority score

Proxies: features naturally correlated with class membership (bias in features)

- unavoidable with rich data

How machines learn to discriminate

Sample size disparity: high samples implies lower error, higher confidence

- **small groups have higher variance**, higher error levels
- in best effort mechanisms, smaller groups ignored to favor majority score

Proxies: features naturally correlated with class membership (bias in features)

- unavoidable with rich data

Example: in unsupervised learning, are features isolating

How machines learn to discriminate

Sample size disparity: high samples implies lower error, higher confidence

- **small groups have higher variance**, higher error levels
- in best effort mechanisms, smaller groups ignored to favor majority score

Proxies: features naturally correlated with class membership (bias in features)

- unavoidable with rich data

Example: in unsupervised learning, are features isolating

- groups of good vs. bad workers?

How machines learn to discriminate

Sample size disparity: high samples implies lower error, higher confidence

- **small groups have higher variance**, higher error levels
- in best effort mechanisms, smaller groups ignored to favor majority score

Proxies: features naturally correlated with class membership (bias in features)

- unavoidable with rich data

Example: in unsupervised learning, are features isolating

- groups of good vs. bad workers?
- whites vs. blacks?

How machines learn to discriminate

Sample size disparity: high samples implies lower error, higher confidence

- **small groups have higher variance**, higher error levels
- in best effort mechanisms, smaller groups ignored to favor majority score

Proxies: features naturally correlated with class membership (bias in features)

- unavoidable with rich data

Example: in unsupervised learning, are features isolating

- groups of good vs. bad workers?
- whites vs. blacks?
- likely a mixture of both (inducing bias)

How machines learn to discriminate

Sample size disparity: high samples implies lower error, higher confidence

- **small groups have higher variance**, higher error levels
- in best effort mechanisms, smaller groups ignored to favor majority score

Proxies: features naturally correlated with class membership (bias in features)

- unavoidable with rich data

Example: in unsupervised learning, are features isolating

- groups of good vs. bad workers?
- whites vs. blacks?
- likely a mixture of both (inducing bias)
- how to enforce orthogonality to unwanted features?

Three different problems to address:

How machines learn to discriminate

Three different problems to address:

1. discovering **unobserved** differences in performance due to skewed/tainted samples

How machines learn to discriminate

Three different problems to address:

1. discovering **unobserved** differences in performance due to skewed/tainted samples
   $\longrightarrow$ Difficult because the data are the "first class citizens": **no access to the genuine data, the ground truth**

How machines learn to discriminate

Three different problems to address:

1. discovering **unobserved** differences in performance due to skewed/tainted samples
   $\longrightarrow$ Difficult because the data are the "first class citizens": **no access to the genuine data, the ground truth**

2. even if data perfect, coping with **observed** differences in performance: sample size disparity, limited features

How machines learn to discriminate

Three different problems to address:

1. discovering **unobserved** differences in performance due to skewed/tainted samples
   $\longrightarrow$ Difficult because the data are the "first class citizens": **no access to the genuine data, the ground truth**

2. even if data perfect, coping with **observed** differences in performance: sample size disparity, limited features

3. understand causes of disparities: identify and eliminate proxies (correlated features).

Outline

Formalizing fairness in machine learning

Formal Setup

Probabilistic setup: (e.g., advertisement display for Software Engineer job position)

Formal Setup

Probabilistic setup: (e.g., advertisement display for Software Engineer job position)

- X: feature vector of an individual

Formal Setup

Probabilistic setup: (e.g., advertisement display for Software Engineer job position)

- X: feature vector of an individual
- $Y \in \{0, 1\}$: target (e.g., bad/good candidate)

Formal Setup

Probabilistic setup: (e.g., advertisement display for Software Engineer job position)

- X: feature vector of an individual
- $Y \in \{0, 1\}$: target (e.g., bad/good candidate)
- A: sensitive attribute (e.g., gender)

Formal Setup

Probabilistic setup: (e.g., advertisement display for Software Engineer job position)

- X: feature vector of an individual
- $Y \in \{0, 1\}$: target (e.g., bad/good candidate)
- A: sensitive attribute (e.g., gender)
- $\hat{Y} = g(X, A) \in \{0, 1\}$: (hard) predictor (e.g., show ad or not)

Formal Setup

Probabilistic setup: (e.g., advertisement display for Software Engineer job position)

- X: feature vector of an individual
- $Y \in \{0, 1\}$: target (e.g., bad/good candidate)
- A: sensitive attribute (e.g., gender)
- $\hat{Y} = g(X, A) \in \{0, 1\}$: (hard) predictor (e.g., show ad or not)
- $R = r(X, A) \in [0, 1]$: (soft) score function (e.g., probability of clicking on ad)

Formal Setup

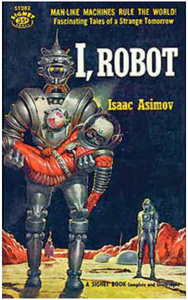
Probabilistic setup: (e.g., advertisement display for Software Engineer job position)

- X: feature vector of an individual
- $Y \in \{0, 1\}$: target (e.g., bad/good candidate)
- A: sensitive attribute (e.g., gender)
- $\hat{Y} = g(X, A) \in \{0, 1\}$: (hard) predictor (e.g., show ad or not)
- $R = r(X, A) \in [0, 1]$: (soft) score function (e.g., probability of clicking on ad)

  $\longrightarrow$ e.g., Bayes' optimal score for quadratic loss (MMSE):

$$R_{\text{Bayes}} = \mathbb{E}[Y|X = x, A = a].$$

The three desiderata

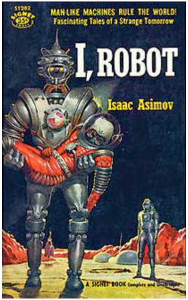


Law 1. **Independence** (also called **demographic parity**)

Law 2. **Separation** (also called **predictive value parity**)

Law 3. **Sufficiency**.

The three desiderata



Law 1. **Independence** (also called **demographic parity**)

Law 2. **Separation** (also called **predictive value parity**)

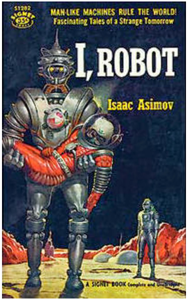
Law 3. **Sufficiency**.

Question: How would the "AI robot" apply the fairness rules?

The three desiderata

Law 1. **Independence**: (also called **demographic parity**)

$$\hat{Y} \perp A$$



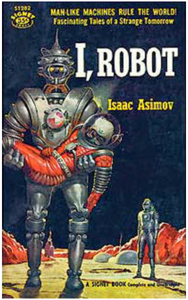
The three desiderata

Law 1. **Independence**: (also called **demographic parity**)

$$\hat{Y} \perp A$$

- equivalently:

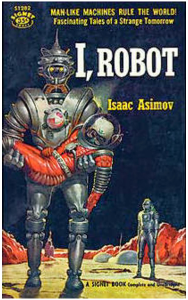
$$\mathbb{P}(\hat{Y} \mid A = a) = \mathbb{P}(\hat{Y} \mid A = b)$$



The three desiderata

Law 1. **Independence**: (also called **demographic parity**)

$$\hat{Y} \perp A$$

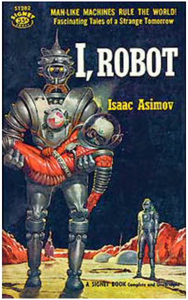


- equivalently:

$$\mathbb{P}(\hat{Y} \mid A = a) = \mathbb{P}(\hat{Y} \mid A = b)$$

$\longrightarrow$ *equal proportion of positive outcomes* ($\hat{Y} = 1$) *in each population*

The three desiderata



Law 1. **Independence**: (also called **demographic parity**)

$$\hat{Y} \perp A$$

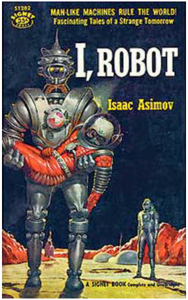
- equivalently:

$$\mathbb{P}(\hat{Y} \mid A = a) = \mathbb{P}(\hat{Y} \mid A = b)$$

$\longrightarrow$ *equal proportion of positive outcomes ($\hat{Y} = 1$) in each population*

- equal average output in each sensitive category

The three desiderata



Law 1. **Independence**: (also called **demographic parity**)

$$\hat{Y} \perp A$$

- equivalently:

$$\mathbb{P}(\hat{Y} \mid A = a) = \mathbb{P}(\hat{Y} \mid A = b)$$

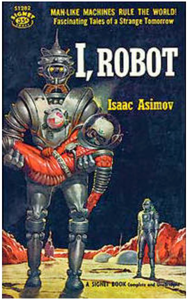
$\longrightarrow$ *equal proportion of positive outcomes* ($\hat{Y} = 1$) *in each population*

- equal average output in each sensitive category
- **Example**: parity in juries, parity in companies (as many women as men)

The three desiderata

Law 1. **Independence**: (also called **demographic parity**)

$$\hat{Y} \perp A$$



- equivalently:

$$\mathbb{P}(\hat{Y} \mid A = a) = \mathbb{P}(\hat{Y} \mid A = b)$$

$\longrightarrow$ *equal proportion of positive outcomes ($\hat{Y} = 1$) in each population*

- equal average output in each sensitive category
- **Example**: parity in juries, parity in companies (as many women as men)

- ε-variants:

$$\frac{\mathbb{P}(\hat{Y} = 1 \mid A = a)}{\mathbb{P}(\hat{Y} = 1 \mid A = b)} \geq 1 - \varepsilon.$$

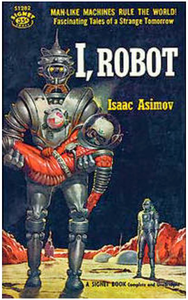
or

$$|\mathbb{P}(\hat{Y} = 1 \mid A = a) - \mathbb{P}(\hat{Y} = 1 \mid A = b)| \leq \varepsilon$$

The three desiderata

Law 1. **Independence**: (also called **demographic parity**)

$$\hat{Y} \perp A$$



- equivalently:

$$\mathbb{P}(\hat{Y} \mid A = a) = \mathbb{P}(\hat{Y} \mid A = b)$$

$\longrightarrow$ *equal proportion of positive outcomes (*$\hat{Y} = 1$*) in each population*

- equal average output in each sensitive category
- **Example**: parity in juries, parity in companies (as many women as men)

- ε-variants:

$$\frac{\mathbb{P}(\hat{Y} = 1 \mid A = a)}{\mathbb{P}(\hat{Y} = 1 \mid A = b)} \geq 1 - \varepsilon.$$

or

$$|\mathbb{P}(\hat{Y} = 1 \mid A = a) - \mathbb{P}(\hat{Y} = 1 \mid A = b)| \leq \varepsilon$$

e.g., the 20% discrimination rule!

The three desiderata

How to achieve independence?:

The three desiderata

How to achieve independence?:

- algorithm postprocessing
- data preprocessing (representation/feature learning)

The three desiderata

How to achieve independence?:

- algorithm postprocessing
- data preprocessing (representation/feature learning)

- e.g., information theory approach

$$Z = \phi(X, A), \quad \text{with} \quad \max I(X; Z) \text{ and } \min I(A; Z)$$

The three desiderata

How to achieve independence?:

- algorithm postprocessing
- data preprocessing (representation/feature learning)

- e.g., information theory approach

$$Z = \phi(X, A), \quad \text{with} \quad \max I(X; Z) \text{ and } \min I(A; Z)$$

then use $\hat{Y} = g(Z, A)$ rather than $\hat{Y} = g(X, A)$.

The three desiderata

How to achieve independence?:

- algorithm postprocessing
- data preprocessing (representation/feature learning)

- e.g., information theory approach

$$Z = \phi(X, A), \quad \text{with} \quad \max I(X; Z) \text{ and } \min I(A; Z)$$

then use $\hat{Y} = g(Z, A)$ rather than $\hat{Y} = g(X, A)$.

Problems:

The three desiderata

How to achieve independence?:

- algorithm postprocessing
- data preprocessing (representation/feature learning)

- e.g., information theory approach

$$Z = \phi(X, A), \quad \text{with} \quad \max I(X; Z) \text{ and } \min I(A; Z)$$

then use $\hat{Y} = g(Z, A)$ rather than $\hat{Y} = g(X, A)$.

Problems:

- ignores possible correlations between Y and A

The three desiderata

How to achieve independence?:

- algorithm postprocessing
- data preprocessing (representation/feature learning)
- e.g., information theory approach

$$Z = \phi(X, A), \quad \text{with} \quad \max I(X; Z) \text{ and } \min I(A; Z)$$

then use $\hat{Y} = g(Z, A)$ rather than $\hat{Y} = g(X, A)$.

Problems:

- ignores possible correlations between Y and A
- **Example**: since more male SWE than female SWE, even with Z independent of A, Y relates highly to A.

The three desiderata

How to achieve independence?:

- algorithm postprocessing
- data preprocessing (representation/feature learning)

- e.g., information theory approach

$$Z = \phi(X, A), \quad \text{with} \quad \max I(X; Z) \text{ and } \min I(A; Z)$$

then use $\hat{Y} = g(Z, A)$ rather than $\hat{Y} = g(X, A)$.

Problems:

- ignores possible correlations between Y and A
- **Example**: since more male SWE than female SWE, even with Z independent of A, Y relates highly to A.

  $\Rightarrow$ Perfect predictor $C = Y$ unreachable.

The three desiderata

How to achieve independence?:

- algorithm postprocessing
- data preprocessing (representation/feature learning)

- e.g., information theory approach

$$Z = \phi(X, A), \quad \text{with} \quad \max I(X; Z) \text{ and } \min I(A; Z)$$

then use $\hat{Y} = g(Z, A)$ rather than $\hat{Y} = g(X, A)$.

Problems:

- ignores possible correlations between Y and A
- **Example**: since more male SWE than female SWE, even with Z independent of A, Y relates highly to A.

  $\Rightarrow$ Perfect predictor $C = Y$ unreachable.

- creates random assignments in one group to avoid discrimination

The three desiderata

How to achieve independence?:

- algorithm postprocessing
- data preprocessing (representation/feature learning)

- e.g., information theory approach

$$Z = \phi(X, A), \quad \text{with} \quad \max I(X; Z) \text{ and } \min I(A; Z)$$

  then use $\hat{Y} = g(Z, A)$ rather than $\hat{Y} = g(X, A)$.

Problems:

- ignores possible correlations between Y and A
- **Example**: since more male SWE than female SWE, even with Z independent of A, Y relates highly to A.

  $\Rightarrow$ Perfect predictor $C = Y$ unreachable.

- creates random assignments in one group to avoid discrimination
  (if for all males, $Y = 0$ (no male candidate suitable), solution is to pick males randomly ($\hat{Y} = 1$) to avoid discrimination!)

The three desiderata

How to achieve independence?:

- algorithm postprocessing
- data preprocessing (representation/feature learning)
- e.g., information theory approach

$$Z = \phi(X, A), \quad \text{with} \quad \max I(X; Z) \text{ and } \min I(A; Z)$$

then use $\hat{Y} = g(Z, A)$ rather than $\hat{Y} = g(X, A)$.

Problems:

- ignores possible correlations between Y and A
- **Example**: since more male SWE than female SWE, even with Z independent of A, Y relates highly to A.

  $\Rightarrow$ Perfect predictor $C = Y$ unreachable.

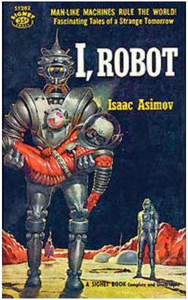
- creates random assignments in one group to avoid discrimination (if for all males, $Y = 0$ (no male candidate suitable), solution is to pick males randomly ($\hat{Y} = 1$) to avoid discrimination!)
- promotes **algorithm laziness!**

The three desiderata

Law 2. **Separation**: (also called **predictive value parity**)

$$R \perp A \mid Y$$

(reminder: $R = r(X, A)$=r(X,A) is the "soft score")



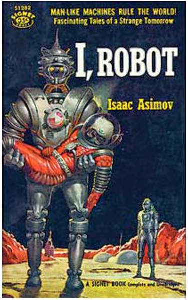
The three desiderata

Law 2. **Separation**: (also called **predictive value parity**)

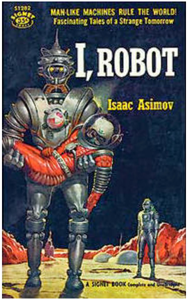
$$R \perp A \mid Y$$

(reminder: $R = r(X, A)$=r(X,A) is the "soft score")

- R and A are independent **conditionally on** Y



The three desiderata



Law 2. **Separation**: (also called **predictive value parity**)

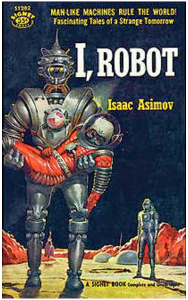
$$R \perp A \mid Y$$

(reminder: $R = r(X, A)$=r(X,A) is the "soft score")

- R and A are independent **conditionally on** Y
- equivalently:

$$\mathbb{P}(R = r \mid Y = y,\ A = a) = \mathbb{P}(R = r \mid Y = y,\ A = b)$$

The three desiderata



Law 2. **Separation**: (also called **predictive value parity**)

$$R \perp A \mid Y$$

(reminder: $R = r(X, A)$=r(X,A) is the "soft score")

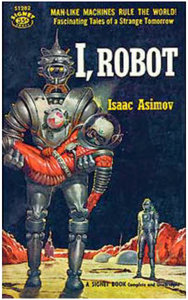
- R and A are independent **conditionally on** Y
- equivalently:

$$\mathbb{P}(R = r \mid Y = y,\ A = a) = \mathbb{P}(R = r \mid Y = y,\ A = b)$$

$\longrightarrow$ *since* $\hat{Y} = \{R > r_0\}$*, equal false positive/negative rates (*$\hat{Y} \neq Y$*) in each population*

$$\mathbb{P}(\hat{Y} = \hat{y} \mid Y = y,\ A = a) = \mathbb{P}(\hat{Y} = \hat{y} \mid Y = y,\ A = b)$$

The three desiderata



Law 2. **Separation**: (also called **predictive value parity**)

$$R \perp A \mid Y$$

(reminder: $R = r(X, A)$=r(X,A) is the "soft score")

- R and A are independent **conditionally on** Y
- equivalently:

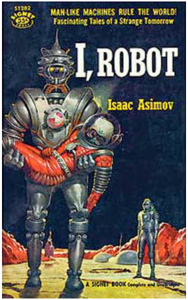
$$\mathbb{P}(R = r \mid Y = y,\ A = a) = \mathbb{P}(R = r \mid Y = y,\ A = b)$$

$\longrightarrow$ *since* $\hat{Y} = \{R > r_0\}$, *equal false positive/negative rates* ($\hat{Y} \neq Y$) *in each population*

$$\mathbb{P}(\hat{Y} = \hat{y} \mid Y = y,\ A = a) = \mathbb{P}(\hat{Y} = \hat{y} \mid Y = y,\ A = b)$$

- in words: equal performance, error rates within each group

The three desiderata



Law 2. Separation: (also called predictive value parity)

$$R \perp A \mid Y$$

(reminder: $R = r(X, A)$=r(X,A) is the "soft score")

- R and A are independent conditionally on Y
- equivalently:

$$\mathbb{P}(R = r \mid Y = y,\ A = a) = \mathbb{P}(R = r \mid Y = y,\ A = b)$$

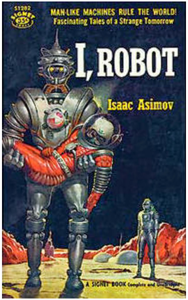
$\longrightarrow$ *since* $\hat{Y} = \{R > r_0\}$, *equal false positive/negative rates* ($\hat{Y} \neq Y$) *in each population*

$$\mathbb{P}(\hat{Y} = \hat{y} \mid Y = y,\ A = a) = \mathbb{P}(\hat{Y} = \hat{y} \mid Y = y,\ A = b)$$

- in words: equal performance, error rates within each group
- graphically:

$$A - Y - R$$

The three desiderata



Law 2. **Separation**: (also called **predictive value parity**)

$$R \perp A \mid Y$$

(reminder: $R = r(X, A)$=r(X,A) is the "soft score")

- R and A are independent **conditionally on** Y
- equivalently:

$$\mathbb{P}(R = r \mid Y = y,\ A = a) = \mathbb{P}(R = r \mid Y = y,\ A = b)$$

$\longrightarrow$ *since* $\hat{Y} = \{R > r_0\}$*, equal false positive/negative rates (*$\hat{Y} \neq Y$*) in each population*

$$\mathbb{P}(\hat{Y} = \hat{y} \mid Y = y,\ A = a) = \mathbb{P}(\hat{Y} = \hat{y} \mid Y = y,\ A = b)$$

- in words: equal performance, error rates within each group
- graphically:

$$A - Y - R$$

(i.e., Y "sits" between A and R.)

The three desiderata

Key properties of separation:

The three desiderata

Key properties of separation:

- compatible with optimality: $R = Y$ allowed.

The three desiderata

Key properties of separation:

- compatible with optimality: $R = Y$ allowed.
- allows A and R to be correlated
- allows A and Y to be correlated

The three desiderata

Key properties of separation:

- compatible with optimality: $R = Y$ allowed.
- allows A and R to be correlated
- allows A and Y to be correlated
- this is fine because A is "confined" in the ground truth Y!

The three desiderata

Key properties of separation:

- compatible with optimality: $R = Y$ allowed.
- allows A and R to be correlated
- allows A and Y to be correlated
- this is fine because A is "confined" in the ground truth Y!
- penalizes laziness: reduces errors **uniformly** on all groups!

The three desiderata

Key properties of separation:

- compatible with optimality: $R = Y$ allowed.
- allows A and R to be correlated
- allows A and Y to be correlated
- this is fine because A is “confined” in the ground truth Y!
- penalizes laziness: reduces errors **uniformly** on all groups!
- if close to optimal unconditionally, still **close to optimal under constraint**,

The three desiderata

Key properties of separation:

- compatible with optimality: $R = Y$ allowed.
- allows A and R to be correlated
- allows A and Y to be correlated
- this is fine because A is “confined” in the ground truth Y!
- penalizes laziness: reduces errors **uniformly** on all groups!
- if close to optimal unconditionally, still **close to optimal under constraint**, i.e.,

$$\mathbb{P}(\hat{Y} = y \mid Y = y) \simeq 1 \Rightarrow \mathbb{P}(\hat{Y} = y \mid Y = y, A = a) \simeq 1$$

The three desiderata

Key properties of separation:

- compatible with optimality: $R = Y$ allowed.
- allows A and R to be correlated
- allows A and Y to be correlated
- this is fine because A is "confined" in the ground truth Y!
- penalizes laziness: reduces errors **uniformly** on all groups!
- if close to optimal unconditionally, still **close to optimal under constraint**, i.e.,

$$\mathbb{P}(\hat{Y} = y \mid Y = y) \simeq 1 \Rightarrow \mathbb{P}(\hat{Y} = y \mid Y = y, A = a) \simeq 1$$

follows from

$$\mathbb{P}(\hat{Y} = y \mid Y = y) = \sum_a \mathbb{P}(\hat{Y} = y \mid Y = y, A = a) \cdot \mathbb{P}(A = a)$$

The three desiderata

Key properties of separation:

- compatible with optimality: $R = Y$ allowed.

- allows A and R to be correlated
- allows A and Y to be correlated
- this is fine because A is "confined" in the ground truth Y!

- penalizes laziness: reduces errors **uniformly** on all groups!

- if close to optimal unconditionally, still **close to optimal under constraint**, i.e.,

$$\mathbb{P}(\hat{Y} = y \mid Y = y) \simeq 1 \Rightarrow \mathbb{P}(\hat{Y} = y \mid Y = y, A = a) \simeq 1$$

  follows from

$$\mathbb{P}(\hat{Y} = y \mid Y = y) = \sum_a \mathbb{P}(\hat{Y} = y \mid Y = y, A = a) \cdot \mathbb{P}(A = a)$$

  so, LHS$\simeq 1 \Rightarrow \mathbb{P}(\hat{Y} = y \mid Y = y, A = a) \simeq 1$ for each a (unless some $\mathbb{P}(A = a) \ll 1$).

The three desiderata

Key properties of separation:

- compatible with optimality: $R = Y$ allowed.
- allows A and R to be correlated
- allows A and Y to be correlated
- this is fine because A is "confined" in the ground truth Y!
- penalizes laziness: reduces errors **uniformly** on all groups!
- if close to optimal unconditionally, still **close to optimal under constraint**, i.e.,

$$\mathbb{P}(\hat{Y} = y \mid Y = y) \simeq 1 \Rightarrow \mathbb{P}(\hat{Y} = y \mid Y = y, A = a) \simeq 1$$

  follows from

$$\mathbb{P}(\hat{Y} = y \mid Y = y) = \sum_a \mathbb{P}(\hat{Y} = y \mid Y = y, A = a) \cdot \mathbb{P}(A = a)$$

  so, LHS$\simeq 1 \Rightarrow \mathbb{P}(\hat{Y} = y \mid Y = y, A = a) \simeq 1$ for each a (unless some $\mathbb{P}(A = a) \ll 1$).

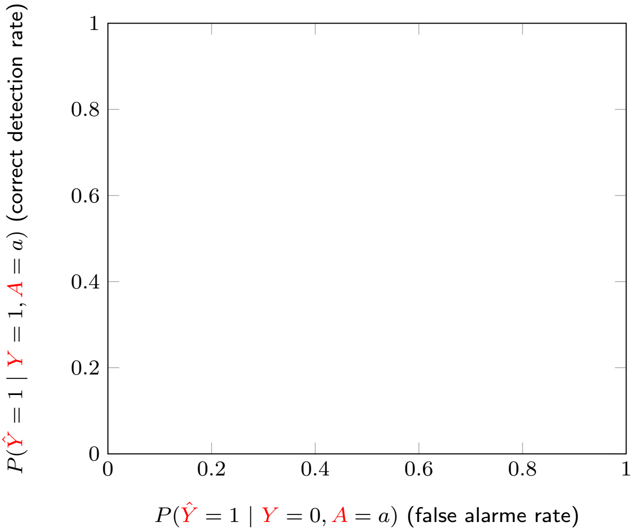
- postprocessing ($R \to \hat{Y}$): any thresholding allowed!

The three desiderata

Postprocessing: ROC curve (receiver operator curve)

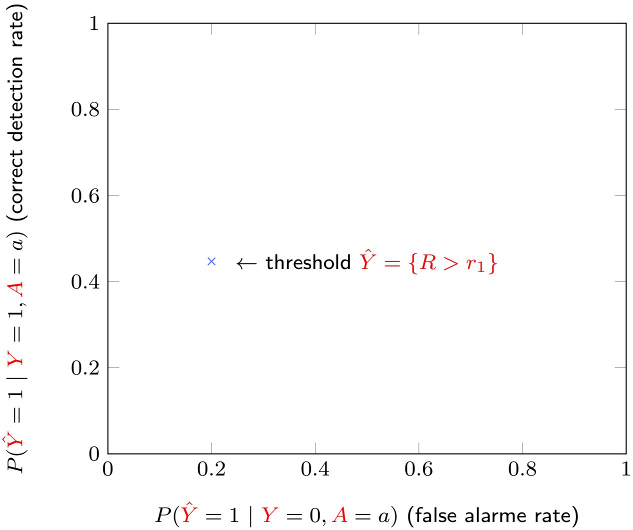
The three desiderata

Postprocessing: ROC curve (receiver operator curve)



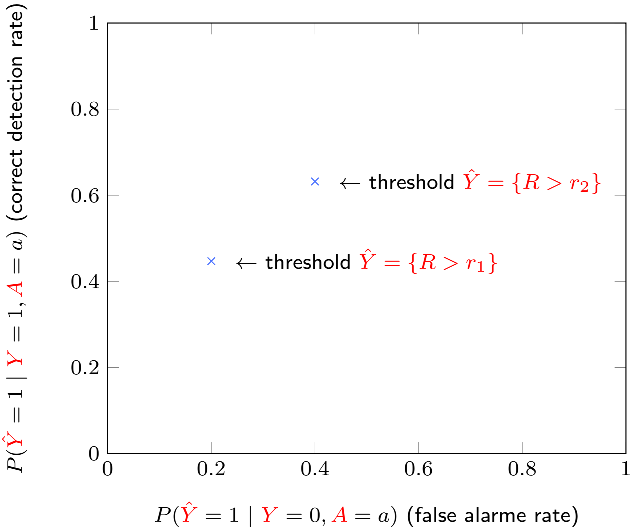
The three desiderata

Postprocessing: ROC curve (receiver operator curve)



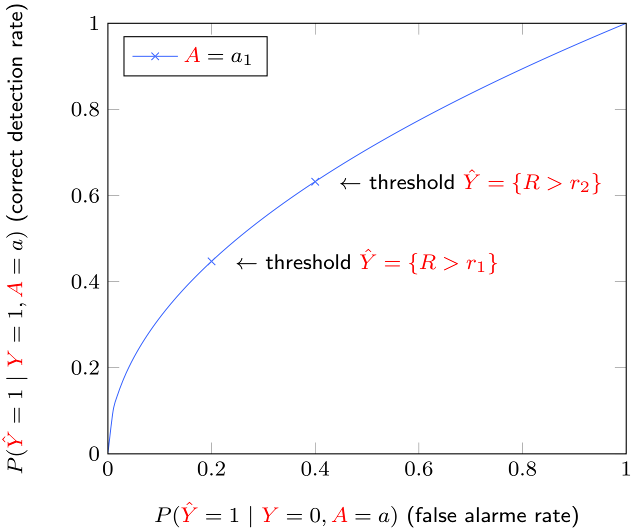
The three desiderata

Postprocessing: ROC curve (receiver operator curve)



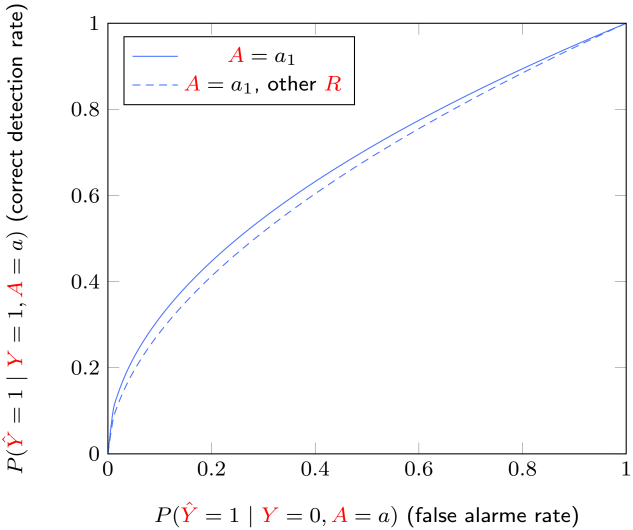
The three desiderata

Postprocessing: ROC curve (receiver operator curve)



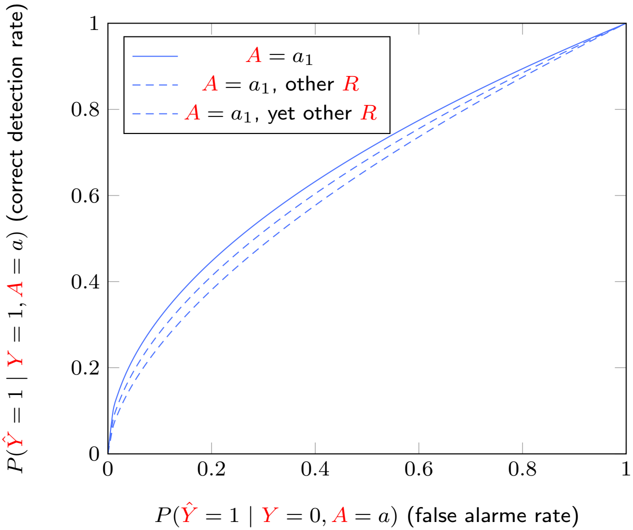
The three desiderata

Postprocessing: ROC curve (receiver operator curve)



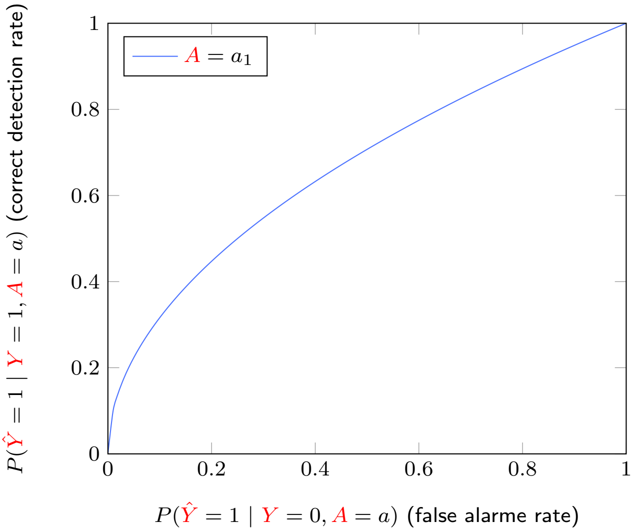
The three desiderata

Postprocessing: ROC curve (receiver operator curve)



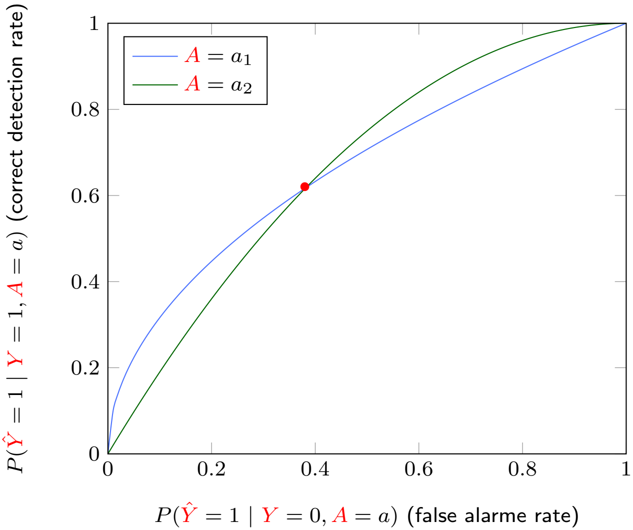
The three desiderata

Postprocessing: ROC curve (receiver operator curve)



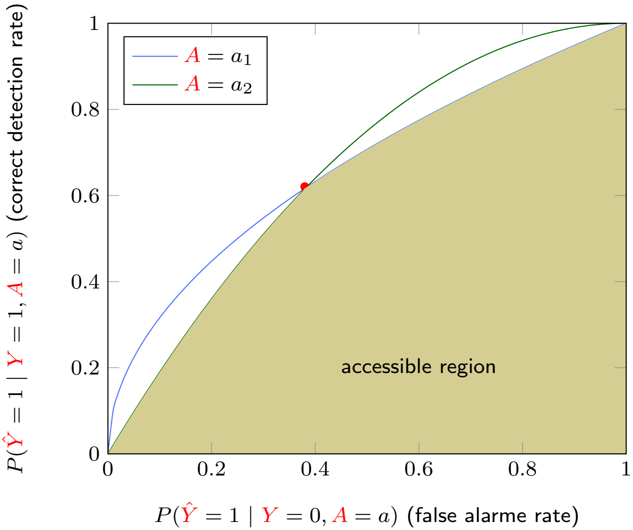
The three desiderata

Postprocessing: ROC curve (receiver operator curve)



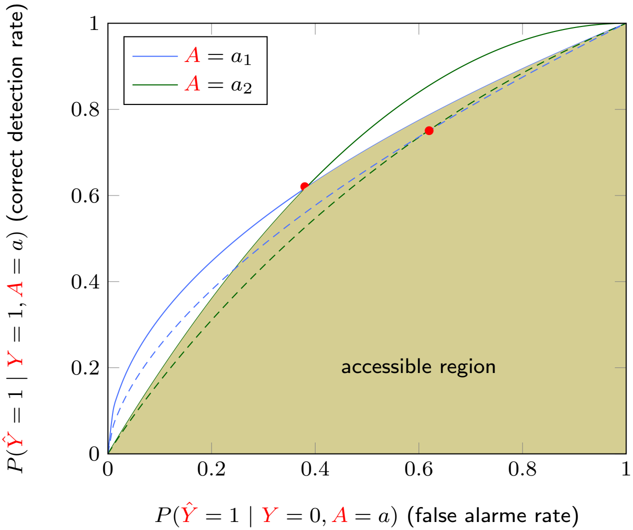
The three desiderata

Postprocessing: ROC curve (receiver operator curve)



The three desiderata

Postprocessing: ROC curve (receiver operator curve)



The three desiderata

Postprocessing: ROC curve (receiver operator curve)

The three desiderata

Postprocessing: ROC curve (receiver operator curve)

- choose decision threshold r such that (recall $R = r(X, A)$)

$$\mathbb{P}(r(X, A = a) > r \mid Y = y, A = a) = \mathbb{P}(r(X, A = b) > r \mid Y = y, A = b)$$

The three desiderata

Postprocessing: ROC curve (receiver operator curve)

- choose decision threshold r such that (recall $R = r(X, A)$)

$$\mathbb{P}(r(X, A = a) > r \mid Y = y, A = a) = \mathbb{P}(r(X, A = b) > r \mid Y = y, A = b)$$

- $\Rightarrow$ crossing point of two conditional decision rules in ROC curve.

The three desiderata

Postprocessing: ROC curve (receiver operator curve)

- choose decision threshold r such that (recall $R = r(X, A)$)

$$\mathbb{P}(r(X, A = a) > r \mid Y = y, A = a) = \mathbb{P}(r(X, A = b) > r \mid Y = y, A = b)$$

- $\Rightarrow$ crossing point of two conditional decision rules in ROC curve.
- **Careful**! Requires score reparametrization **or** different thresholds $R > r_a \mid A = a$.

The three desiderata

Postprocessing: ROC curve (receiver operator curve)

- choose decision threshold r such that (recall $R = r(X, A)$)

$$\mathbb{P}(r(X, A = a) > r \mid Y = y, A = a) = \mathbb{P}(r(X, A = b) > r \mid Y = y, A = b)$$

- $\Rightarrow$ crossing point of two conditional decision rules in ROC curve.
- **Careful**! Requires score reparametrization **or** different thresholds $R > r_a \mid A = a$.

Reparametrization: assume two intersecting ROC curves

$$\begin{aligned} f_a(r) &= (x_a(r), y_a(r)) = (\mathrm{FAR}_a(r), \mathrm{CDR}_a(r)) \quad \text{for} \quad r = r(X, A = a) \\ f_b(r) &= (x_b(r), y_b(r)) = (\mathrm{FAR}_b(r), \mathrm{CDR}_b(r)) \quad \text{for} \quad r = r(X, A = b) \end{aligned}$$

The three desiderata

Postprocessing: ROC curve (receiver operator curve)

- choose decision threshold r such that (recall $R = r(X, A)$)

$$\mathbb{P}(r(X, A = a) > r \mid Y = y, A = a) = \mathbb{P}(r(X, A = b) > r \mid Y = y, A = b)$$

- $\Rightarrow$ crossing point of two conditional decision rules in ROC curve.
- **Careful**! Requires score reparametrization **or** different thresholds $R > r_a \mid A = a$.

Reparametrization: assume two intersecting ROC curves

$$f_a(r) = (x_a(r), y_a(r)) = (\mathrm{FAR}_a(r), \mathrm{CDR}_a(r)) \quad \text{for} \quad r = r(X, A = a)$$
$$f_b(r) = (x_b(r), y_b(r)) = (\mathrm{FAR}_b(r), \mathrm{CDR}_b(r)) \quad \text{for} \quad r = r(X, A = b)$$

(in particular, $f.(0) = 0$, $f.(1) = 1$)

The three desiderata

Postprocessing: ROC curve (receiver operator curve)

- choose decision threshold r such that (recall $R = r(X, A)$)

$$\mathbb{P}(r(X, A = a) > r \mid Y = y, A = a) = \mathbb{P}(r(X, A = b) > r \mid Y = y, A = b)$$

- $\Rightarrow$ crossing point of two conditional decision rules in ROC curve.
- **Careful**! Requires score reparametrization **or** different thresholds $R > r_a \mid A = a$.

Reparametrization: assume two intersecting ROC curves

$$f_a(r) = (x_a(r), y_a(r)) = (\mathrm{FAR}_a(r), \mathrm{CDR}_a(r)) \quad \text{for} \quad r = r(X, A = a)$$
$$f_b(r) = (x_b(r), y_b(r)) = (\mathrm{FAR}_b(r), \mathrm{CDR}_b(r)) \quad \text{for} \quad r = r(X, A = b)$$

(in particular, $f_\cdot(0) = 0$, $f_\cdot(1) = 1$)

- intersection defined as

$$f_a(r_1) = f_b(r_2) \quad \text{for some} \quad r_1, r_2.$$

The three desiderata

Postprocessing: ROC curve (receiver operator curve)

- choose decision threshold r such that (recall $R = r(X, A)$)

$$\mathbb{P}(r(X, A = a) > r \mid Y = y, A = a) = \mathbb{P}(r(X, A = b) > r \mid Y = y, A = b)$$

- $\Rightarrow$ crossing point of two conditional decision rules in ROC curve.
- **Careful**! Requires score reparametrization **or** different thresholds $R > r_a \mid A = a$.

Reparametrization: assume two intersecting ROC curves

$$f_a(r) = (x_a(r), y_a(r)) = (\mathrm{FAR}_a(r), \mathrm{CDR}_a(r)) \quad \text{for} \quad r = r(X, A = a)$$
$$f_b(r) = (x_b(r), y_b(r)) = (\mathrm{FAR}_b(r), \mathrm{CDR}_b(r)) \quad \text{for} \quad r = r(X, A = b)$$

(in particular, $f_\cdot(0) = 0$, $f_\cdot(1) = 1$)

- intersection defined as

$$f_a(r_1) = f_b(r_2) \quad \text{for some} \quad r_1, r_2.$$

- **Unlikely that $r_1 = r_2$!** Depends on parametrization.

The three desiderata

Postprocessing: ROC curve (receiver operator curve)

- choose decision threshold r such that (recall $R = r(X, A)$)

$$\mathbb{P}(r(X, A = a) > r \mid Y = y, A = a) = \mathbb{P}(r(X, A = b) > r \mid Y = y, A = b)$$

- $\Rightarrow$ crossing point of two conditional decision rules in ROC curve.
- **Careful**! Requires score reparametrization **or** different thresholds $R > r_a \mid A = a$.

Reparametrization: assume two intersecting ROC curves

$$f_a(r) = (x_a(r), y_a(r)) = (\mathrm{FAR}_a(r), \mathrm{CDR}_a(r)) \quad \text{for} \quad r = r(X, A = a)$$
$$f_b(r) = (x_b(r), y_b(r)) = (\mathrm{FAR}_b(r), \mathrm{CDR}_b(r)) \quad \text{for} \quad r = r(X, A = b)$$

(in particular, $f_\cdot(0) = 0$, $f_\cdot(1) = 1$)

- intersection defined as

$$f_a(r_1) = f_b(r_2) \quad \text{for some} \quad r_1, r_2.$$

- **Unlikely that $r_1 = r_2$!** Depends on parametrization.
- Reparametrization: When intersecting couple (r_1, r_2) found, **scale parameters $r \to r' = h(r)$ so that $f_a \to f'_a$, $f_b \to f'_b$** and

$$f'_a(r) = f_a(h(r_1)) = f_a(r_1) = f_b(r_2) = f_b(h_b(r)) = f'_b(r).$$

The three desiderata

Alternatives to postprocessing:

The three desiderata

Alternatives to postprocessing:

- collect more data (to improve ROC curves $\Rightarrow$ both curves will tend to merge)

The three desiderata

Alternatives to postprocessing:

- collect more data (to improve ROC curves $\Rightarrow$ both curves will tend to merge)
- achieve constraint at training time: solve

$$\min_g \mathbb{E}[\ell(r(X, A), Y)]$$

such that $r(X, A) \perp A \mid Y$

The three desiderata

Alternatives to postprocessing:

- collect more data (to improve ROC curves $\Rightarrow$ both curves will tend to merge)
- achieve constraint at training time: solve

$$\min_g \mathbb{E}[\ell(r(X,A),Y)]$$

such that $r(X,A) \perp A \mid Y$

- generically intractable!

The three desiderata

Alternatives to postprocessing:

- collect more data (to improve ROC curves $\Rightarrow$ both curves will tend to merge)
- achieve constraint at training time: solve

$$\min_g \mathbb{E}[\ell(r(X,A),Y)]$$

  such that $r(X,A) \perp A \mid Y$

- generically intractable!

- doable in **joint Gaussian** case (vector (A,Y,R)) with quadratic loss:

The three desiderata

Alternatives to postprocessing:

- collect more data (to improve ROC curves $\Rightarrow$ both curves will tend to merge)
- achieve constraint at training time: solve

$$\min_g \mathbb{E}[\ell(r(X, A), Y)]$$

  such that $r(X, A) \perp A \mid Y$

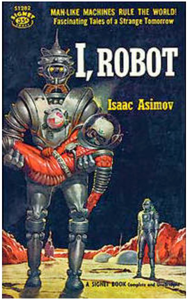
- generically intractable!
- doable in **joint Gaussian** case (vector (A, Y, R)) with quadratic loss: equivalent to imposing

$$\sigma_{RA}\sigma_Y^2 = \sigma_{RY}\sigma_{YA}.$$

The three desiderata

Law 3. **Sufficiency**:

$$Y \perp A \mid R$$

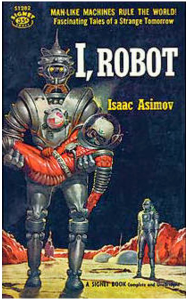


The three desiderata

Law 3. **Sufficiency**:

$$Y \perp A \mid R$$

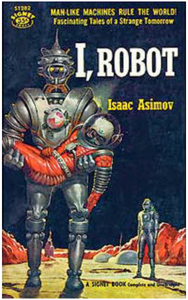
- Y and A are independent **conditionally on** R



The three desiderata

Law 3. **Sufficiency**:

$$Y \perp A \mid R$$



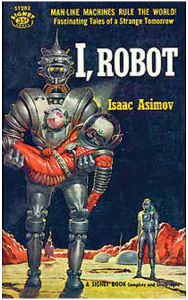
- Y and A are independent **conditionally on** R
- equivalently

$$\mathbb{P}(Y = y \mid R = r, A = a) = \mathbb{P}(Y = y \mid R = r, A = b)$$

The three desiderata

Law 3. **Sufficiency**:

$$Y \perp A \mid R$$

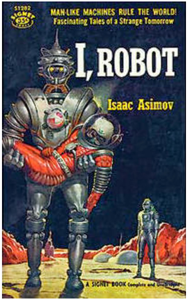


- Y and A are independent **conditionally on** R
- equivalently

$$\mathbb{P}(Y = y \mid R = r, A = a) = \mathbb{P}(Y = y \mid R = r, A = b)$$

$\longrightarrow$ *if* $\hat{Y} = R \in \{0, 1\}$, *equal genuine positive/negative rates in selected population*

The three desiderata



Law 3. **Sufficiency**:

$$Y \perp A \mid R$$

- Y and A are independent **conditionally on** R
- equivalently

$$\mathbb{P}(Y = y \mid R = r, A = a) = \mathbb{P}(Y = y \mid R = r, A = b)$$

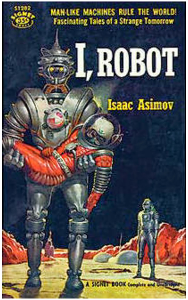
$\longrightarrow$ *if* $\hat{Y} = R \in \{0,1\}$, *equal genuine positive/negative rates in selected population*

- in words: R is **sufficient** to establish Y (**and** A)

The three desiderata

Law 3. **Sufficiency**:

$$Y \perp A \mid R$$



- Y and A are independent **conditionally on** R
- equivalently

$$\mathbb{P}(Y = y \mid R = r, A = a) = \mathbb{P}(Y = y \mid R = r, A = b)$$

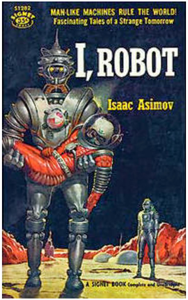
$\longrightarrow$ *if $\hat{Y} = R \in \{0, 1\}$, equal genuine positive/negative rates in selected population*

- in words: R is **sufficient** to establish Y (**and** A)
- or: for the purpose of predicting Y, no need to see A when we have R
  (R is **sufficient** to predict Y, no need to look at A)

The three desiderata

Law 3. **Sufficiency**:

$$Y \perp A \mid R$$



- Y and A are independent **conditionally on** R
- equivalently

$$\mathbb{P}(Y = y \mid R = r, A = a) = \mathbb{P}(Y = y \mid R = r, A = b)$$

$\longrightarrow$ *if* $\hat{Y} = R \in \{0, 1\}$, *equal genuine positive/negative rates in selected population*

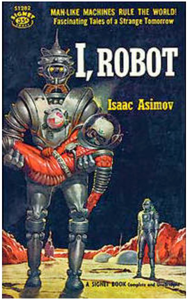
- in words: R is **sufficient** to establish Y (**and** A)
- or: for the purpose of predicting Y, no need to see A when we have R
  (R is **sufficient** to predict Y, no need to look at A)
- graphically:

$$A - R - Y$$

The three desiderata

Law 3. **Sufficiency**:

$$Y \perp A \mid R$$



- Y and A are independent **conditionally on** R
- equivalently

$$\mathbb{P}(Y = y \mid R = r, A = a) = \mathbb{P}(Y = y \mid R = r, A = b)$$

$\longrightarrow$ *if* $\hat{Y} = R \in \{0, 1\}$, *equal genuine positive/negative rates in selected population*

- in words: R is **sufficient** to establish Y (**and** A)
- or: for the purpose of predicting Y, no need to see A when we have R
  (R is **sufficient** to predict Y, no need to look at A)

- graphically:

$$A - R - Y$$

(i.e., Y "sits" between A and R.)

The three desiderata

Properties of sufficiency:

The three desiderata

Properties of sufficiency:

- why is it desirable?

The three desiderata

Properties of sufficiency:

- why is it desirable?
- *example*: for credit allocation decision, no need to look at gender, race when making decision: **the score is sufficient!**

The three desiderata

Properties of sufficiency:

- why is it desirable?
- *example*: for credit allocation decision, no need to look at gender, race when making decision: **the score is sufficient!**
  ($\Rightarrow$ Good for legal matters)

The three desiderata

Properties of sufficiency:

- why is it desirable?
- *example*: for credit allocation decision, no need to look at gender, race when making decision: **the score is sufficient!**
  ($\Rightarrow$ Good for legal matters)

  Careful!: but the score $R = r(X, A)$ would likely **depend indirectly on** race, gender!

The three desiderata

Properties of sufficiency:

- why is it desirable?
- *example*: for credit allocation decision, no need to look at gender, race when making decision: **the score is sufficient!**
  ($\Rightarrow$ Good for legal matters)

  Careful!: but the score $R = r(X, A)$ would likely **depend indirectly on** race, gender!

- sufficiency implied by **group-wise calibration**:

$$\mathbb{P}(Y = 1 \mid R = r, A = a) = r.$$

The three desiderata

Group-wise calibration: Platt scaling to obtain

$$\mathbb{P}(Y = 1 \mid R = r, A = a) = r.$$

The three desiderata

Group-wise calibration: Platt scaling to obtain

$$\mathbb{P}(Y = 1 \mid R = r, A = a) = r.$$

- for uncalibrated R, **fit R to a sigmoid**

$$S = \frac{1}{1 + \exp(\alpha R + \beta)}$$

The three desiderata

Group-wise calibration: Platt scaling to obtain

$$\mathbb{P}(Y = 1 \mid R = r, A = a) = r.$$

- for uncalibrated R, **fit R to a sigmoid**

$$S = \frac{1}{1 + \exp(\alpha R + \beta)}$$

in such a way to minimize the cross-entropy loss

$$-\mathbb{E}\left[Y \log S + (1 - Y) \log(1 - S)\right]$$

The three desiderata

Group-wise calibration: Platt scaling to obtain

$$\mathbb{P}(Y = 1 \mid R = r, A = a) = r.$$

- for uncalibrated R, **fit R to a sigmoid**

$$S = \frac{1}{1 + \exp(\alpha R + \beta)}$$

in such a way to minimize the cross-entropy loss

$$-\mathbb{E}\left[Y \log S + (1 - Y)\log(1 - S)\right]$$

i.e., minimize KL-divergence $\mathrm{KL}(Y; S)$.

The three desiderata

Group-wise calibration: Platt scaling to obtain

$$\mathbb{P}(Y = 1 \mid R = r, A = a) = r.$$

- for uncalibrated R, **fit R to a sigmoid**

$$S = \frac{1}{1 + \exp(\alpha R + \beta)}$$

  in such a way to minimize the cross-entropy loss

$$-\mathbb{E}\left[Y \log S + (1 - Y) \log(1 - S)\right]$$

  i.e., minimize KL-divergence $\mathrm{KL}(Y; S)$.

- this enforces

$$\mathbb{P}(Y = 1 \mid S = s, A = a) \simeq s.$$

The three desiderata

Group-wise calibration: Platt scaling to obtain

$$\mathbb{P}(Y = 1 \mid R = r, A = a) = r.$$

- for uncalibrated R, **fit R to a sigmoid**

$$S = \frac{1}{1 + \exp(\alpha R + \beta)}$$

  in such a way to minimize the cross-entropy loss

$$-\mathbb{E}\left[Y \log S + (1 - Y) \log(1 - S)\right]$$

  i.e., minimize KL-divergence $\mathrm{KL}(Y; S)$.
- this enforces

$$\mathbb{P}(Y = 1 \mid S = s, A = a) \simeq s.$$

- set decision threshold

$$S > \frac{1}{2} \Rightarrow \hat{Y} = 1$$

The three desiderata

Group-wise calibration: Platt scaling to obtain

$$\mathbb{P}(Y = 1 \mid R = r, A = a) = r.$$

- for uncalibrated R, **fit R to a sigmoid**

$$S = \frac{1}{1 + \exp(\alpha R + \beta)}$$

  in such a way to minimize the cross-entropy loss

$$-\mathbb{E}\left[Y \log S + (1 - Y) \log(1 - S)\right]$$

  i.e., minimize KL-divergence $\mathrm{KL}(Y; S)$.
- this enforces

$$\mathbb{P}(Y = 1 \mid S = s, A = a) \simeq s.$$

- set decision threshold

$$S > \frac{1}{2} \Rightarrow \hat{Y} = 1$$

- since cross-entropy loss unknown, calibration performed on **training dataset** $\{(y_i, r_i)\}_{i=1}^n$:

$$\min_{\alpha,\beta} - \sum_{i=1}^{n} y_i \log s_i + (1 - y_i) \log(1 - s_i) \quad \text{where} \quad s_i = \frac{1}{1 + \exp(\alpha r_i + \beta)}.$$

The three desiderata

MAJOR PROBLEM

The three desiderata

MAJOR PROBLEM

Any two of the **3 desiderata are mutually exclusive**! (except in trivial cases)

The three desiderata

MAJOR PROBLEM

Any two of the **3 desiderata are mutually exclusive**! (except in trivial cases)

Consequences:

The three desiderata

MAJOR PROBLEM

Any two of the **3 desiderata are mutually exclusive**! (except in trivial cases)

Consequences:

- in practice, **trade-offs** must be performed

The three desiderata

MAJOR PROBLEM

Any two of the **3 desiderata are mutually exclusive**! (except in trivial cases)

Consequences:

- in practice, **trade-offs** must be performed
- this explains **(theoretically!)** why lawsuits can be endless!

The three desiderata

MAJOR PROBLEM

Any two of the **3 desiderata are mutually exclusive**! (except in trivial cases)

Consequences:

- in practice, **trade-offs** must be performed
- this explains **(theoretically!)** why lawsuits can be endless!
- which optimal balancing of desiderata for each given situation, ML problem?

MAJOR PROBLEM

Any two of the **3 desiderata are mutually exclusive**! (except in trivial cases)

Consequences:

- in practice, **trade-offs** must be performed
- this explains **(theoretically!)** why lawsuits can be endless!
- which optimal balancing of desiderata for each given situation, ML problem?
- more philosophically: **is fairness accessible to mathematics, and thus machines?**

The three desiderata

Independence vs. sufficiency:

The three desiderata

Independence vs. sufficiency:

Proposition

If $Y \not\perp A$, then **either independence holds or sufficiency, but not both**.

The three desiderata

Independence vs. sufficiency:

Proposition

If $Y \not\perp A$, then **either independence holds or sufficiency, but not both**.

Proof

If $Y \not\perp A$ (non trivial case) and $Y \perp A \mid R$ (sufficiency), then $R \not\perp A$ (no independence).

The three desiderata

Independence vs. sufficiency:

Proposition

If $Y \not\perp A$, then **either independence holds or sufficiency, but not both**.

Proof

If $Y \not\perp A$ (non trivial case) and $Y \perp A \mid R$ (sufficiency), then $R \not\perp A$ (no independence).

So, conversely, if $R \perp A$ (independence), then $Y \not\perp A \mid R$ (not sufficiency) or $Y \perp A$ (trivial case).

The three desiderata

Independence vs. separation:

The three desiderata

Independence vs. separation:

Proposition

If $Y \not\perp A$ and $Y \not\perp R$, then **either independence holds or separation, but not both**.

The three desiderata

Independence vs. separation:

Proposition

If $Y \not\perp A$ and $Y \not\perp R$, then **either independence holds or separation, but not both**.

Proof

If $R \perp A$ and $R \perp A \mid Y$, then $A \perp Y$ or $R \perp Y$.

The three desiderata

Independence vs. separation:

Proposition

If $Y \not\perp A$ and $Y \not\perp R$, then **either independence holds or separation, but not both**.

Proof

If $R \perp A$ and $R \perp A \mid Y$, then $A \perp Y$ or $R \perp Y$.

So, conversely, if $A \not\perp Y$ and $R \not\perp Y$, then either $R \not\perp A$ (not independence) or $R \not\perp A \mid Y$ (not separation).

The three desiderata

Separation vs. sufficiency:

The three desiderata

Separation vs. sufficiency:

Proposition

Assume all events in (A, R, Y) have positive probability. Then, if $A \not\perp Y$, **either separation or sufficiency holds, but not both**.

The three desiderata

Separation vs. sufficiency:

Proposition

Assume all events in (A, R, Y) have positive probability. Then, if $A \not\perp Y$, **either separation or sufficiency holds, but not both**.

Proof

It can be shown that $A \perp R \mid Y$ and $A \perp Y \mid R$ implies $A \perp (R, Y)$ (which implies $A \perp Y$).

The three desiderata

Separation vs. sufficiency:

Proposition

Assume all events in (A, R, Y) have positive probability. Then, if $A \not\perp Y$, **either separation or sufficiency holds, but not both**.

Proof

It can be shown that $A \perp R \mid Y$ and $A \perp Y \mid R$ implies $A \perp (R, Y)$ (which implies $A \perp Y$).

Hence, $A \not\perp Y$ implies either $A \not\perp R \mid Y$ or $A \not\perp Y \mid R$.

The three desiderata

Separation vs. sufficiency:

Proposition

Assume all events in (A, R, Y) have positive probability. Then, if $A \not\perp Y$, **either separation or sufficiency holds, but not both**.

Proof

It can be shown that $A \perp R \mid Y$ and $A \perp Y \mid R$ implies $A \perp (R, Y)$ (which implies $A \perp Y$).

Hence, $A \not\perp Y$ implies either $A \not\perp R \mid Y$ or $A \not\perp Y \mid R$.

So, conversely, separation **and** sufficiency imply $A \perp Y$ which is forbidden (trivial setting).

Outline

Case study: loan granting

Loan granting: the setup

Borrowed from:

https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Setup:

Loan granting: the setup

Borrowed from:

https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Setup:

- 2 sensitive populations: **blue** and **orange** (variable A)

Loan granting: the setup

Borrowed from:

https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Setup:

- 2 sensitive populations: **blue** and **orange** (variable A)
- loan decision: $\hat{Y} = \{R > r_0\}$ with
  - $R =$"credit score" (evaluated likelihood to pay back) (based on income, situation, age, etc: possibly correlated to **color**.)

Loan granting: the setup

Borrowed from:

https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Setup:

- 2 sensitive populations: **blue** and **orange** (variable A)
- loan decision: $\hat{Y} = \{R > r_0\}$ with
  - $R =$"credit score" (evaluated likelihood to pay back) (based on income, situation, age, etc: possibly correlated to **color**.)
  - $r_0 =$"loan threshold"

Loan granting: the setup

Borrowed from:

https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Setup:

- 2 sensitive populations: **blue** and **orange** (variable A)
- loan decision: $\hat{Y} = \{R > r_0\}$ with
  - $R =$"credit score" (evaluated likelihood to pay back) (based on income, situation, age, etc: possibly correlated to **color**.)
  - $r_0 =$"loan threshold"
  - $\hat{Y} \in \{0, 1\} =$"gets the loan or not"

Loan granting: the setup

Borrowed from:

https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Setup:

- 2 sensitive populations: **blue** and **orange** (variable A)
- loan decision: $\hat{Y} = \{R > r_0\}$ with
  - $R =$"credit score" (evaluated likelihood to pay back) (based on income, situation, age, etc: possibly correlated to **color**.)
  - $r_0 =$"loan threshold"
  - $\hat{Y} \in \{0, 1\} =$"gets the loan or not"
- expected output $Y =$"will pay back".

Loan granting: the setup

Borrowed from:

https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Setup:

- 2 sensitive populations: **blue** and **orange** (variable A)
- loan decision: $\hat{Y} = \{R > r_0\}$ with
  - $R =$"credit score" (evaluated likelihood to pay back) (based on income, situation, age, etc: possibly correlated to **color**.)
  - $r_0 =$"loan threshold"
  - $\hat{Y} \in \{0, 1\} =$"gets the loan or not"
- expected output $Y =$"will pay back".

Output for the bank:

Loan granting: the setup

Borrowed from:

https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Setup:

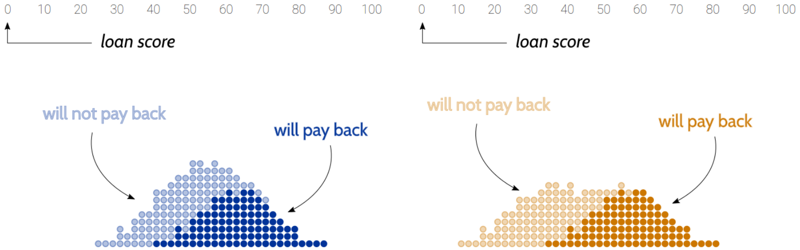
- 2 sensitive populations: **blue** and **orange** (variable A)
- loan decision: $\hat{Y} = \{R > r_0\}$ with
  - $R =$"credit score" (evaluated likelihood to pay back) (based on income, situation, age, etc: possibly correlated to **color**.)
  - $r_0 =$"loan threshold"
  - $\hat{Y} \in \{0, 1\} =$"gets the loan or not"
- expected output $Y =$"will pay back".

Output for the bank:

- successful loan: $\$300$,
- unsuccessful loan: $-\$700$,
- credit score in $(0, 100)$.

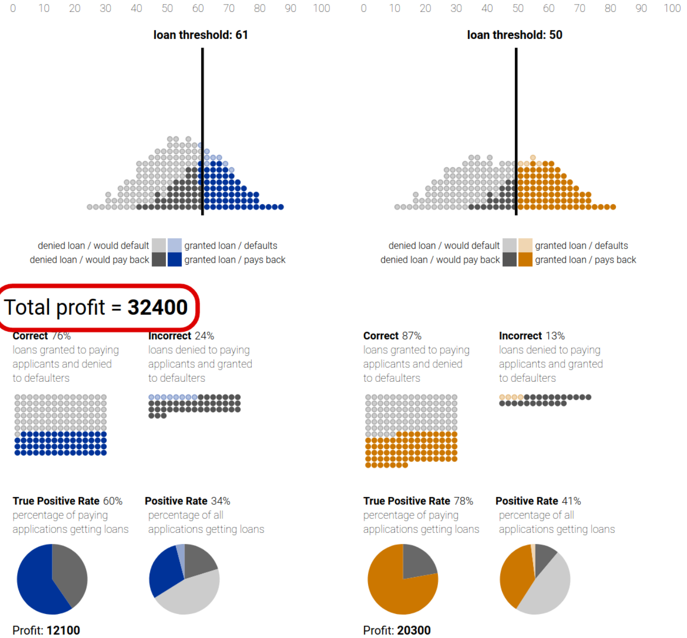
Loan granting: the setup

Populations and credit score:



Loan granting: Max profit

No fairness case: max profit for bank (assuming bank knows statistics)



Loan granting: Max profit

No fairness case: max profit for bank (assuming bank knows statistics)

Loan granting: Max profit

No fairness case: max profit for bank (assuming bank knows statistics)

Discussion:

Loan granting: Max profit

No fairness case: max profit for bank (assuming bank knows statistics)

Discussion:

- highly unfair according to all rules!

Loan granting: Max profit

No fairness case: max profit for bank (assuming bank knows statistics)

Discussion:

- highly unfair according to all rules!
- disparate positive rates $\hat{Y} \mid A$ (34% vs. 41%)

Loan granting: Max profit

No fairness case: max profit for bank (assuming bank knows statistics)

Discussion:

- highly unfair according to all rules!
- disparate positive rates $\hat{Y} \mid A$ (34% vs. 41%)
  $\Rightarrow$ **No demographic parity**

Loan granting: Max profit

No fairness case: max profit for bank (assuming bank knows statistics)

Discussion:

- highly unfair according to all rules!
- disparate positive rates $\hat{Y} \mid A$ (34% vs. 41%)
  $\Rightarrow$ **No demographic parity**
- disparate true positives $\mathbb{P}(\hat{Y} = 1 \mid Y = 1, A = a)$ (60% vs. 78%)

Loan granting: Max profit

No fairness case: max profit for bank (assuming bank knows statistics)

Discussion:

- highly unfair according to all rules!
- disparate positive rates $\hat{Y} \mid A$ (34% vs. 41%)
  $\Rightarrow$ **No demographic parity**
- disparate true positives $\mathbb{P}(\hat{Y} = 1 \mid Y = 1, A = a)$ (60% vs. 78%)
  $\Rightarrow$ **No predictive value parity**

Loan granting: Max profit

No fairness case: max profit for bank (assuming bank knows statistics)

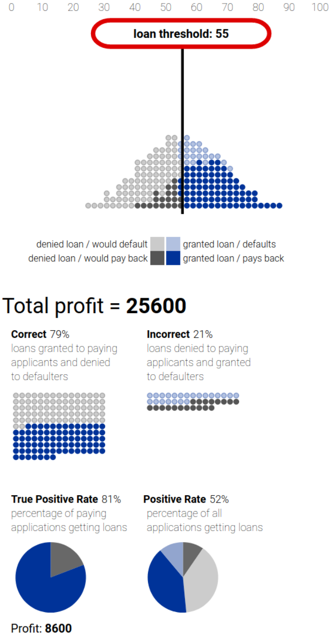
Discussion:

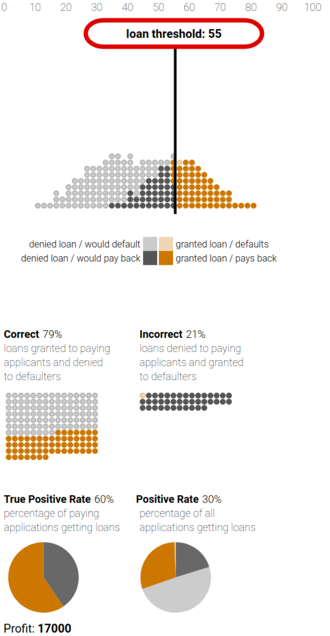
- highly unfair according to all rules!
- disparate positive rates $\hat{Y} \mid A$ (34% vs. 41%)
  $\Rightarrow$ **No demographic parity**
- disparate true positives $\mathbb{P}(\hat{Y} = 1 \mid Y = 1, A = a)$ (60% vs. 78%)
  $\Rightarrow$ **No predictive value parity**

"**The most profitable, since there are no constraints**"

Loan granting: Group unaware

Group unaware case: max profit by considering all groups as one (unique threshold r_0)





Loan granting: Group unaware

Group unaware case: max profit by considering all groups as one (unique threshold r_0)

Loan granting: Group unaware

Group unaware case: max profit by considering all groups as one (unique threshold r_0)

Discussion:

Loan granting: Group unaware

Group unaware case: max profit by considering all groups as one (unique threshold r_0)

Discussion:

- again, highly unfair according to all rules!

Loan granting: Group unaware

Group unaware case: max profit by considering all groups as one (unique threshold r_0)

Discussion:

- again, highly unfair according to all rules!
- disparate positive rates $\hat{Y} \mid A$ (52% vs. 30%)

Loan granting: Group unaware

Group unaware case: max profit by considering all groups as one (unique threshold r_0)

Discussion:

- again, highly unfair according to all rules!
- disparate positive rates $\hat{Y} \mid A$ (52% vs. 30%)
  $\Rightarrow$ **No demographic parity**

Loan granting: Group unaware

Group unaware case: max profit by considering all groups as one (unique threshold r_0)

Discussion:

- again, highly unfair according to all rules!
- disparate positive rates $\hat{Y} \mid A$ (52% vs. 30%)
  $\Rightarrow$ **No demographic parity**
- disparate true positives $\mathbb{P}(\hat{Y} = 1 \mid Y = 1, A = a)$ (81% vs. 60%)

Loan granting: Group unaware

Group unaware case: max profit by considering all groups as one (unique threshold r_0)

Discussion:

- again, highly unfair according to all rules!
- disparate positive rates $\hat{Y} \mid A$ (52% vs. 30%)
  $\Rightarrow$ **No demographic parity**
- disparate true positives $\mathbb{P}(\hat{Y} = 1 \mid Y = 1, A = a)$ (81% vs. 60%)
  $\Rightarrow$ **No predictive value parity**

Loan granting: Group unaware

Group unaware case: max profit by considering all groups as one (unique threshold r_0)

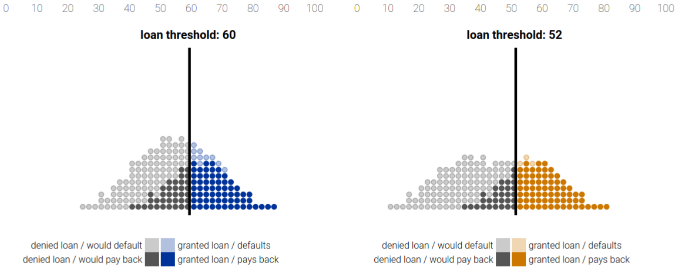
Discussion:

- again, highly unfair according to all rules!
- disparate positive rates $\hat{Y} \mid A$ (52% vs. 30%)
  $\Rightarrow$ **No demographic parity**
- disparate true positives $\mathbb{P}(\hat{Y} = 1 \mid Y = 1, A = a)$ (81% vs. 60%)
  $\Rightarrow$ **No predictive value parity**

"**Both groups have the same threshold**"

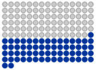
Loan granting: Demographic parity

Demographic parity case: Independence $\hat{Y} \perp A$ (law 1)



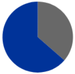
Total profit = **30800**

Correct 77%
loans granted to paying applicants and denied to defaulters



Incorrect 23%
loans denied to paying applicants and granted to defaulters

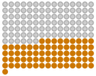
True Positive Rate 64%
percentage of paying applications getting loans



Positive Rate 37%
percentage of all applications getting loans

Profit: **11900**

Correct 84%
loans granted to paying applicants and denied to defaulters



Incorrect 16%
loans denied to paying applicants and granted to defaulters

True Positive Rate 71%
percentage of paying applications getting loans

Positive Rate 37%
percentage of all applications getting loans

Profit: **18900**

Loan granting: Demographic parity

Demographic parity case: Independence $\hat{Y} \perp A$ (law 1)

Loan granting: Demographic parity

Demographic parity case: Independence $\hat{Y} \perp A$ (law 1)

Discussion:

Loan granting: Demographic parity

Demographic parity case: Independence $\hat{Y} \perp A$ (law 1)

Discussion:

- demographic fairness: equal outputs in each population (disregarding worth)

Loan granting: Demographic parity

Demographic parity case: Independence $\hat{Y} \perp A$ (law 1)

Discussion:

- demographic fairness: equal outputs in each population (disregarding worth)
- equal positive rates $\hat{Y} \mid A$ (37% vs. 37%)

Loan granting: Demographic parity

Demographic parity case: Independence $\hat{Y} \perp A$ (law 1)

Discussion:

- demographic fairness: equal outputs in each population (disregarding worth)
- equal positive rates $\hat{Y} \mid A$ (37% vs. 37%)
  $\Rightarrow$ **Demographic parity enforced!**

Loan granting: Demographic parity

Demographic parity case: Independence $\hat{Y} \perp A$ (law 1)

Discussion:

- demographic fairness: equal outputs in each population (disregarding worth)
- equal positive rates $\hat{Y} \mid A$ (37% vs. 37%)
  $\Rightarrow$ **Demographic parity enforced!**
- disparate true positives $\mathbb{P}(\hat{Y} = 1 \mid Y = 1, A = a)$ (64% vs. 71%)

Loan granting: Demographic parity

Demographic parity case: Independence $\hat{Y} \perp A$ (law 1)

Discussion:

- demographic fairness: equal outputs in each population (disregarding worth)
- equal positive rates $\hat{Y} \mid A$ (37% vs. 37%)
  $\Rightarrow$ **Demographic parity enforced!**
- disparate true positives $\mathbb{P}(\hat{Y} = 1 \mid Y = 1, A = a)$ (64% vs. 71%)
  $\Rightarrow$ **No predictive value parity**

Loan granting: Demographic parity

Demographic parity case: Independence $\hat{Y} \perp A$ (law 1)

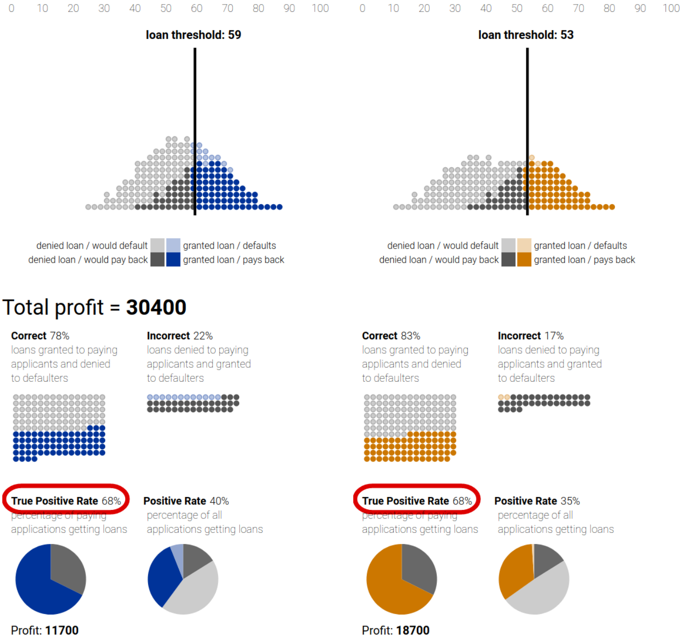
Discussion:

- demographic fairness: equal outputs in each population (disregarding worth)
- equal positive rates $\hat{Y} \mid A$ (37% vs. 37%)
  $\Rightarrow$ **Demographic parity enforced!**
- disparate true positives $\mathbb{P}(\hat{Y} = 1 \mid Y = 1, A = a)$ (64% vs. 71%)
  $\Rightarrow$ **No predictive value parity**

"**The number of loans given to each group is the same**"

Loan granting: Equal opportunity

Equal opportunity case: Separation $R \perp A \mid Y$ (law 2)



Loan granting: Equal opportunity

Equal opportunity case: Separation $R \perp A \mid Y$ (law 2)

Loan granting: Equal opportunity

Equal opportunity case: Separation $R \perp A \mid Y$ (law 2)

Discussion:

Loan granting: Equal opportunity

Equal opportunity case: Separation $R \perp A \mid Y$ (law 2)

Discussion:

- **equal worth**: same opportunities in subpopulations

Loan granting: Equal opportunity

Equal opportunity case: Separation $R \perp A \mid Y$ (law 2)

Discussion:

- **equal worth**: same opportunities in subpopulations
- disparate positive rates $\hat{Y} \mid A$ (40% vs. 35%)

Loan granting: Equal opportunity

Equal opportunity case: Separation $R \perp A \mid Y$ (law 2)

Discussion:

- **equal worth**: same opportunities in subpopulations
- disparate positive rates $\hat{Y} \mid A$ (40% vs. 35%)
  $\Rightarrow$ **No demographic parity**

Loan granting: Equal opportunity

Equal opportunity case: Separation $R \perp A \mid Y$ (law 2)

Discussion:

- **equal worth**: same opportunities in subpopulations
- disparate positive rates $\hat{Y} \mid A$ (40% vs. 35%)
  $\Rightarrow$ **No demographic parity**
- equal true positives $\mathbb{P}(\hat{Y} = 1 \mid Y = 1, A = a)$ (68% vs. 68%)

Loan granting: Equal opportunity

Equal opportunity case: Separation $R \perp A \mid Y$ (law 2)

Discussion:

- **equal worth**: same opportunities in subpopulations
- disparate positive rates $\hat{Y} \mid A$ (40% vs. 35%)
  $\Rightarrow$ **No demographic parity**
- equal true positives $\mathbb{P}(\hat{Y} = 1 \mid Y = 1, A = a)$ (68% vs. 68%)
  $\Rightarrow$ **Predictive value parity enforced!**

Loan granting: Equal opportunity

Equal opportunity case: Separation $R \perp A \mid Y$ (law 2)

Discussion:

- **equal worth**: same opportunities in subpopulations
- disparate positive rates $\hat{Y} \mid A$ (40% vs. 35%)
  $\Rightarrow$ **No demographic parity**
- equal true positives $\mathbb{P}(\hat{Y} = 1 \mid Y = 1, A = a)$ (68% vs. 68%)
  $\Rightarrow$ **Predictive value parity enforced!**

"**Among people who would pay back a loan, blue and orange groups do equally well**"

Outline

Conclusion. . . well, partial!

Conclusion: fairness in AI

Not a classical course: (you may have noticed!)

Conclusion: fairness in AI

Not a classical course: (you may have noticed!)

- mixing concepts of **ethics, law, and mathematical formalism**

Conclusion: fairness in AI

Not a classical course: (you may have noticed!)

- mixing concepts of **ethics, law, and mathematical formalism**
- very young field, **few conceptual and formal developments**

Conclusion: fairness in AI

Not a classical course: (you may have noticed!)

- mixing concepts of **ethics, law, and mathematical formalism**
- very young field, **few conceptual and formal developments**
- a lot of problems!: **self-contradicting rules**

Conclusion: fairness in AI

Not a classical course: (you may have noticed!)

- mixing concepts of **ethics, law, and mathematical formalism**
- very young field, **few conceptual and formal developments**
- a lot of problems!: **self-contradicting rules**
- but also an unavoidable field in the future of AI!

Conclusion: fairness in AI

Not a classical course: (you may have noticed!)

- mixing concepts of **ethics, law, and mathematical formalism**
- very young field, **few conceptual and formal developments**
- a lot of problems!: **self-contradicting rules**
- but also an unavoidable field in the future of AI!

For more information and developments:

Conclusion: fairness in AI

Not a classical course: (you may have noticed!)

- mixing concepts of **ethics, law, and mathematical formalism**
- very young field, **few conceptual and formal developments**
- a lot of problems!: **self-contradicting rules**
- but also an unavoidable field in the future of AI!

For more information and developments:

- Fairness in ML book: https://fairmlbook.org/
- Video tutorial: https://fairmlbook.org/tutorial1.html
- Google "attacking discrimination in ML" highlight: https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Conclusion: fairness in AI

Not a classical course: (you may have noticed!)

- mixing concepts of **ethics, law, and mathematical formalism**
- very young field, **few conceptual and formal developments**
- a lot of problems!: **self-contradicting rules**
- but also an unavoidable field in the future of AI!

For more information and developments:

- Fairness in ML book: https://fairmlbook.org/
- Video tutorial: https://fairmlbook.org/tutorial1.html
- Google "attacking discrimination in ML" highlight: https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Final thoughts:

Conclusion: fairness in AI

Not a classical course: (you may have noticed!)

- mixing concepts of **ethics, law, and mathematical formalism**
- very young field, **few conceptual and formal developments**
- a lot of problems!: **self-contradicting rules**
- but also an unavoidable field in the future of AI!

For more information and developments:

- Fairness in ML book: https://fairmlbook.org/
- Video tutorial: https://fairmlbook.org/tutorial1.html
- Google "attacking discrimination in ML" highlight: https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Final thoughts:

mathematicians used to be physicists and **philosophers** *until each field got too complex*

Conclusion: fairness in AI

Not a classical course: (you may have noticed!)

- mixing concepts of **ethics, law, and mathematical formalism**
- very young field, **few conceptual and formal developments**
- a lot of problems!: **self-contradicting rules**
- but also an unavoidable field in the future of AI!

For more information and developments:

- Fairness in ML book: https://fairmlbook.org/
- Video tutorial: https://fairmlbook.org/tutorial1.html
- Google "attacking discrimination in ML" highlight: https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Final thoughts:

mathematicians used to be physicists and **philosophers** *until each field got too complex*
what about AI and ethics? should we (as AI experts) become philosophers again?