AI Ethics for AI Practitioners

A design framework for building towards algorithmic justice

Willie Costello

The hazards of algorithmic systems

Bias

Discrimination

Racism

Sexism

Surveillance

MACHINE BIAS

The Tiger Mom Tax: Asians Are Nearly Twice as Likely to Get a Higher Price from Princeton Review

by Julia Angwin, Surya Mattu and Jeff Larson, Sept. 1, 2015, 10 a.m. EDT

Internet Culture

Google Maps' White House glitch, Flickr auto-tag, and the case of the racist algorithm

HIDDEN BIAS

By Claire Cain Miller

July 9, 2015

147

BUSINESS

When Discrimination Is Baked Into Algorithms

As more companies and services use data to target individuals, those analytics could inadvertently amplify bias.

LAUREN KIRCHNER SEPTEMBER 6, 2015

Flickr faces complaints over 'offensive' auto-tagging for photos

Auto-tagging system slaps 'animal' and 'ape' labels on images of black people, and tags concentration camps with 'jungle gym' and 'sport'

ProPublica

Donate

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

The ideals of algorithmic systems

Fairness

Responsibility

Transparency

Accountability

Privacy

The AI ethics ecosystem

Where do we need AI ethics to happen?

- Product level
- Executive level
- Industry level
- Governmental level
- Research level
- Engineering level → AI & ML practitioners

About me

Willie Costello
Data scientist, PhD Philosophy

williecostello.com
linkedin.com/in/williecostello
@williecostello

Ethics is no fuzzier than data science!

Designing (functionally) good algorithms

Crucial design questions

- Do we optimize for accuracy, sensitivity, specificity, something else?
- Do we optimize for predictive power or interpretability?
- How do we ensure there aren't any blindspots in the training data?

There are no universal answers to these questions! The answers in any particular situation will depend on that algorithm's use case

Designing (ethically) good algorithms

Crucial design questions

- ???
- ???
- ???

There are no universal answers to these questions! The answers in any particular situation will depend on that algorithm's use case

The structure of the framework

1 basic question, asked across 3 components and 3 levels of algorithmic actions

	3 components of algorithmic actions		
3 levels	?	?	?
	?	?	?
	?	?	?

The result: 9 specific questions to ask of any algorithmic system

The framework's fundamental ethical concept:

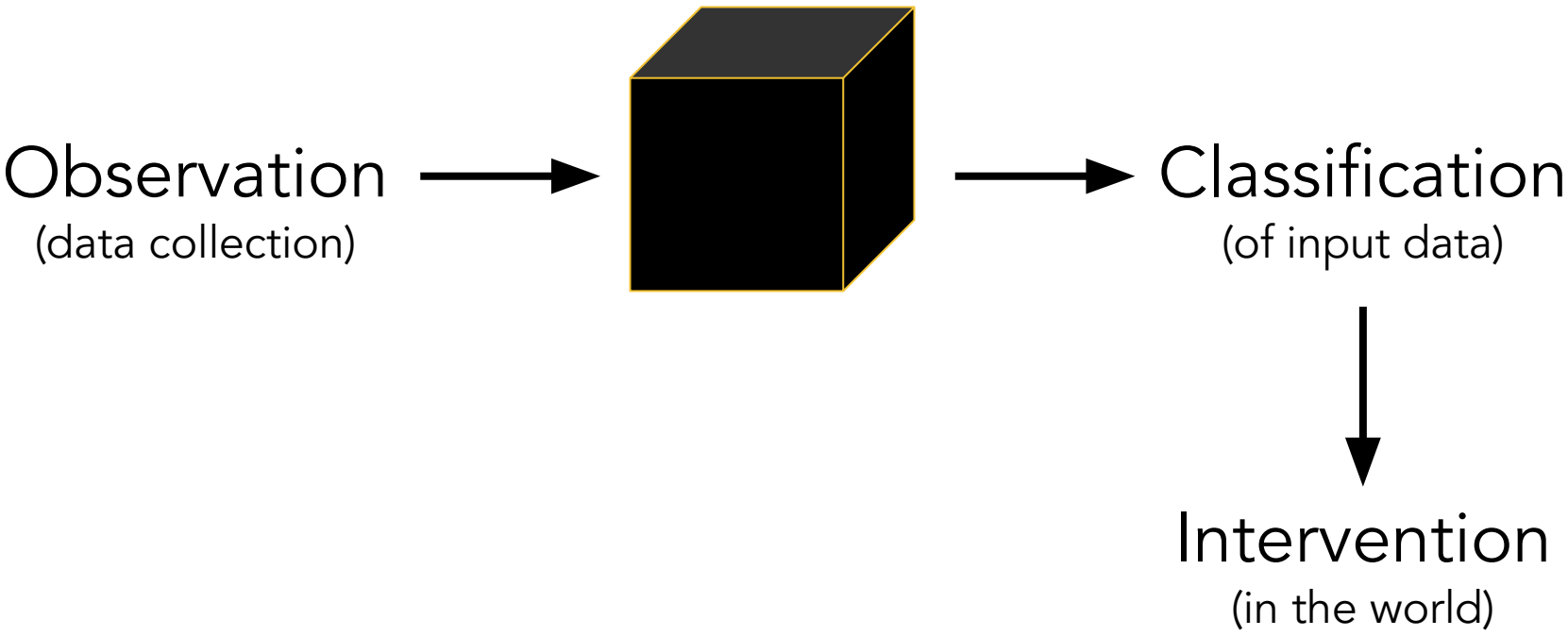
Respect for persons

- Always treat individuals as autonomous human beings
- Treat others always as an end and never simply as a means
- Refrain from manipulating others, violating their rights, and interfering with their own decision-making and self-governance

The framework's central question:

Does this action
respect the persons affected by it?

Three components of algorithmic actions



Three levels of algorithmic actions

Individual

(how an action affects a single person)

Collective

(how an action affects a population of persons)

Iterative

(how an action affects persons when repeated and reiterated)

A matrix of algorithmic actions

Level \ Component	Observation	Classification	Intervention
Individual			
Collective			
Iterative			

A matrix of algorithmic ethics

Level \ Component	Observation	Classification	Intervention
Individual	Does this action respect persons?	Does this action respect persons?	Does this action respect persons?
Collective	Does this action respect persons?	Does this action respect persons?	Does this action respect persons?
Iterative	Does this action respect persons?	Does this action respect persons?	Does this action respect persons?

Respect with observations

The example: loan approval algorithm

Observation action: collecting applicant’s credit history, salary, browsing history

Questions of respect

- Overarching: Are people able to be aware of the data the algorithm is collecting?
- Individual: Does the data collection process respect each individual's rights (e.g., to privacy)?
- Collective: Does the data collection process treat all individuals fairly and equally?
- Iterative: Does the data collection process allow individuals to take an active role in shaping their data?

Respect with classifications

The example: pre-trial risk assessment algorithm

Classification action: classifying detainee as “high risk” of reoffending

Questions of respect

- Overarching: Are people able to dispute the classification the algorithm makes?
- Individual: Does the classification process respect each individual's rights (e.g., to non-discrimination)?
- Collective: Does the classification process treat all individuals fairly and equally?
- Iterative: Does the classification process allow individuals to take an active role in shaping their classification?

Respect with interventions

The example: targeted advertising algorithm

Intervention action: showing an ad to an user according to their “user profile”

Questions

- Overarching: Are people able to act freely in response to the intervention the algorithm makes?
- Individual: Does the intervention respect each individual's rights (e.g., to personal liberty)?
- Collective: Does the intervention treat all individuals fairly and equally?
- Iterative: Does the intervention allow individuals to take an active role in shaping the intervention they experience?

The 6 specific questions of respect

Components of algorithmic actions

- Observation: Are people able to be aware of the data the algorithm is collecting?
- Classification: Are people able to dispute the classification the algorithm makes?
- Intervention: Are people able to act freely in response to the algorithm's intervention?

Levels of algorithmic actions

- Individual: Does the action respect each individual's rights?
- Collective: Does the action treat all individuals fairly and equally?
- Iterative: Does the action allow individuals to take an active role in shaping the action?

The ideals that respect promotes

Components of algorithmic actions

- Observation: transparency
- Classification: voice
- Intervention: autonomy

Levels of algorithmic actions

- Individual: rights
- Collective: fairness
- Iterative: autonomy

Putting the framework into practice

Level \ Component	Observation Are people able to be aware of the data the algorithm is collecting?	Classification Are people able to dispute the classification the algorithm makes?	Intervention Are people able to act freely in response to the algorithm's intervention?
Individual Does the action respect each individual's rights?			
Collective Does the action treat all individuals fairly and equally?			
Iterative Does the action allow individuals to take an active role in shaping the action?			

Level \ Component	Observation Are people able to be aware of the data the algorithm is collecting?	Classification Are people able to dispute the classification the algorithm makes?	Intervention Are people able to act freely in response to the algorithm's intervention?
Individual Does the action respect each individual's rights?	👍		
Collective Does the action treat all individuals fairly and equally?			
Iterative Does the action allow individuals to take an active role in shaping the action?			

Level \ Component	Observation Are people able to be aware of the data the algorithm is collecting?	Classification Are people able to dispute the classification the algorithm makes?	Intervention Are people able to act freely in response to the algorithm's intervention?
Individual Does the action respect each individual's rights?	👍		
Collective Does the action treat all individuals fairly and equally?	👎		
Iterative Does the action allow individuals to take an active role in shaping the action?			

Level \ Component	Observation Are people able to be aware of the data the algorithm is collecting?	Classification Are people able to dispute the classification the algorithm makes?	Intervention Are people able to act freely in response to the algorithm's intervention?
Individual Does the action respect each individual's rights?	👍	👎	👎
Collective Does the action treat all individuals fairly and equally?	👎	👍	👍
Iterative Does the action allow individuals to take an active role in shaping the action?	👍	👍	👎

Level \ Component	Observation Are people able to be aware of the data the algorithm is collecting?	Classification Are people able to dispute the classification the algorithm makes?	Intervention Are people able to act freely in response to the algorithm's intervention?
Individual Does the action respect each individual's rights?		👎	👎
Collective Does the action treat all individuals fairly and equally?	👎		
Iterative Does the action allow individuals to take an active role in shaping the action?			👎

Putting the framework into practice

- Analyze your algorithm into its component actions
- Ask the questions of respect
- Identify problematic actions
- Formulate responses to these actions
- Incorporate responses into your algorithm's design
- Communicate your responses

Thank you!

For a copy of these slides, go to williecostello.com/aiethics

Follow me on Twitter @williecostello
and on LinkedIn at linkedin.com/in/williecostello