

Value-Sensitive Design in Machine Learning: Fairness and Data Subjects

Ethics Lecture for DS4400, Spring 2022

Vance Ricks

Associate Teaching Professor
of Philosophy and Computer
Science

Department of
Philosophy/Religion,
Northeastern University

v.ricks@northeastern.edu

Agenda for the Week

Two days ago

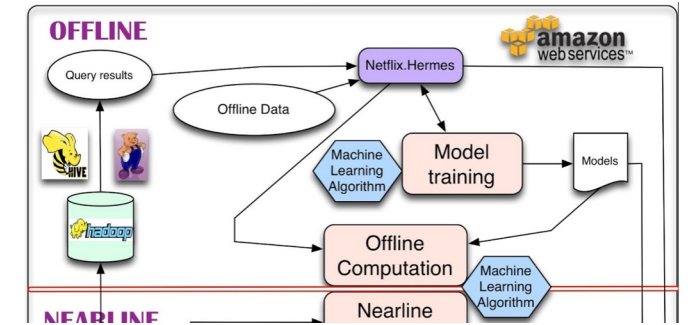
- Introduction: machine learning algorithms in the wild
- SOTBF: using a simulation to uncover ethical questions
- Three conceptions of “fairness” and “unfairness”
- KHASM: Treating people as data subjects

Today

- Quick review of Monday’s materials
- Articulating values and identifying stakeholders: using value-sensitive design (VSD)
- Revisiting SOTBF and KHASM
- The Crisis Text Line, Loris.ai, and the impossibility of informed consent (?)
- Conclusion: Keeping the human in machine learning

Machine Learning Based Computer Aided Diagnosis of Breast Cancer Utilizing Anthropometric and Clinical Features

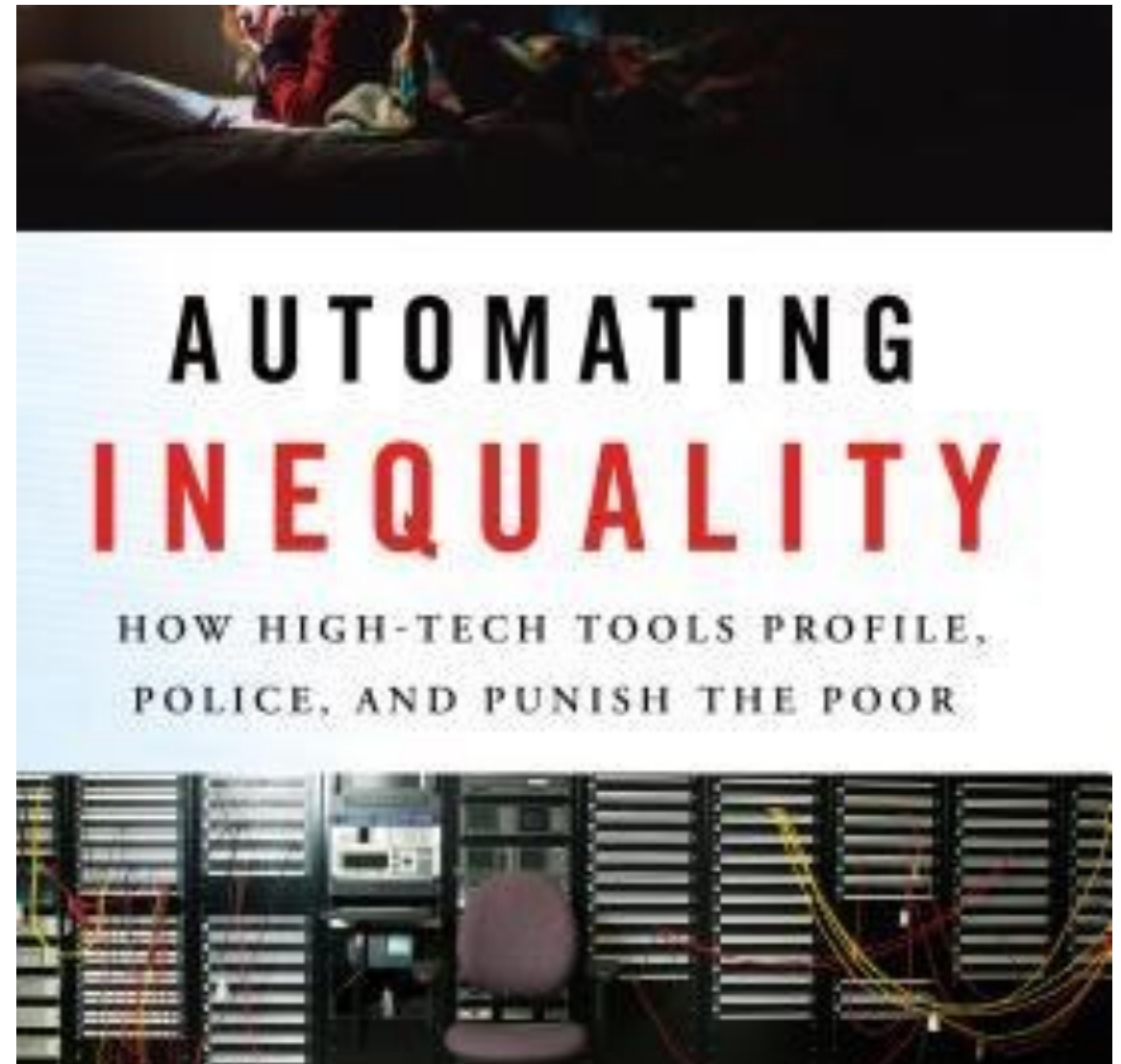
How To Design A Spam Filtering System with Machine Learning Algorithm



Machine Learning Algorithms In the Wild

The Family and Social Services Administration (FSSA) of Indiana provides welfare, food stamps, public health insurance

- goals defined as to reduce fraud, spending and number of those on welfare
- prior to automation, FSSA erred on side of providing benefits: False Pos rate = 4.4% False Neg rate = 1.5%
- after automation, erred on opposite side: FP rate = 6.2% FN rate = 12.2%
- when denied, no explanation given for why
- did not use records from previous system, requiring all new applications



An illustration of a person with dark hair, wearing a light blue shirt and dark pants, standing on the left side of the frame. To their right is a vertical list of five topics, each preceded by a grey circle. Further right is a red door with a sign above it that says 'OFFICE'. The background is a light beige color with a red horizontal line at the top and a white horizontal line at the bottom.

Fairness(es)

Bias(es)

Training data collection practices

“Comprehensibility” of the
algorithms

People as data subjects

OFFICE

Discussion of SOBTF results

Some review questions:

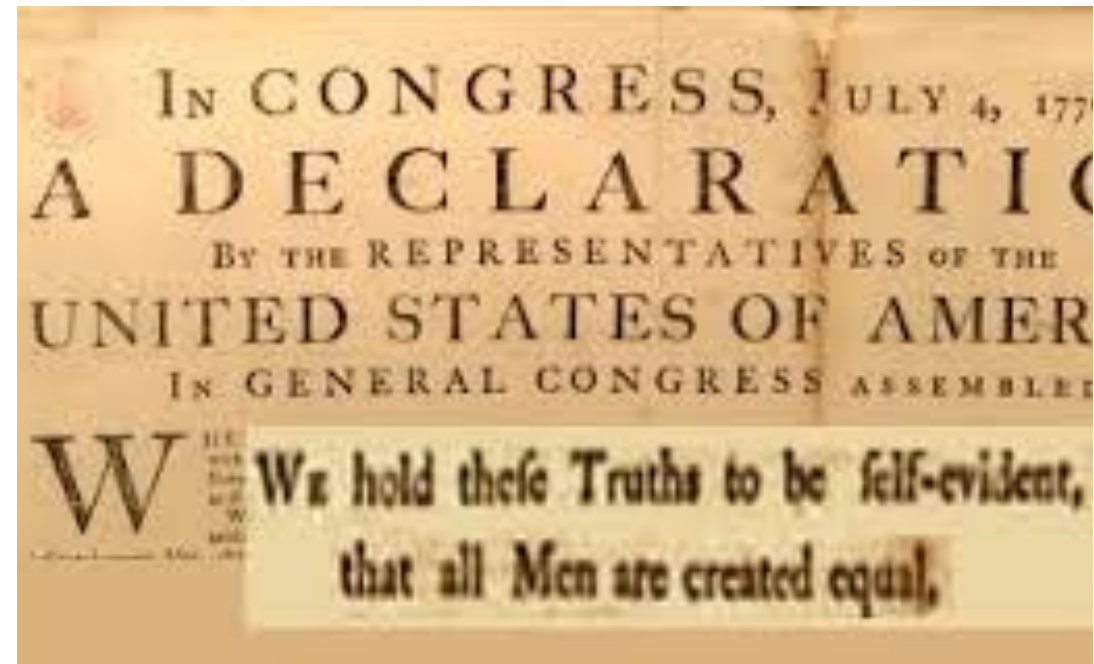
- What does it mean to treat people fairly?
- What are the three main ways that algorithms that automate decision-making might treat people unfairly?
- Describe some of the ways that training data might be corrupted or biased, and result in unfair treatment.

Three conceptions of “fairness” and “unfairness”

1. Fair treatment as a MORAL norm: *people* are treated fairly when those who are similarly situated are treated similarly (when like cases are treated alike).

Any decision to treat *classes* of persons differently should be rationally related to achieving the purpose for which the classification is made. That principle reflects the moral equality of persons.

For example, suppose Prof. Oprea decides to give “A”s to everyone with curly hair and give everyone else “F”s. This is unfair to **both** groups of students, because your hair texture isn’t relevant to what your grade in this course should be.



2. “Fair treatment” as a legal norm

“No state shall make or enforce any law which shall abridge the privileges or immunities of citizens of the United States; nor shall any state deprive any person of life, liberty, or property, without due process of law; nor deny to any person within its jurisdiction the **equal protection** of the laws” (14th Amendment).



3. Fairness as a *distributive* norm:

“patterned” views

v

“process/
procedural” views

Patterned views

Given some goods that we want to distribute,

A **distribution** of those goods is fair if and only if it distributes them in accordance with some (morally acceptable) **property or pattern**

(every *citizen of the US* has the same set of Constitutionally-guaranteed rights)

Process/Procedural views

Given some goods that we want to distribute,

A **distribution** of those goods is fair if and only if it results from some (morally justifiable) process or procedure

(e.g., everyone *who is randomly chosen* to receive free healthcare for life, gets free healthcare for life)

Cases of unfairness in ML algorithms

Here are three ways that ML algorithms that automate decision making may treat people unfairly:

1) **In their purpose (goals):** the algorithm is designed to achieve a goal that is *itself* illegitimate, because that goal relies on false assumptions or reinforces attitudes or patterns of unjustified inequality

(continued) Here are three ways that ML algorithms that automate decision making may treat people unfairly:

2) **In their data collection practices (training data):** the algorithm is not as *accurate* as it could be because of poorly chosen target variables, underlying bias reproduced in training examples, unrepresentative samples, or coarse features

3) **In their distribution of burdens of error (outcomes):** the data and algorithm are as good as possible, but the algorithm imposes greater burdens of error on some stakeholders than others, often in ways that reinforce existing patterns of inequality in society

1. In Purposes (bad or flawed goal)

For all these reasons, there's a growing recognition among scholars and advocates that some biased AI systems should not be "fixed," but abandoned. As co-author Meredith Whittaker said, "We need to look beyond technical fixes for social problems. We need to ask: Who has power? Who is harmed? Who benefits? And ultimately, who gets to decide how these tools are built and which purposes they serve?"

"It's not biased" ≠ "It's morally harmless"

From Vox, "Some AI just shouldn't exist", 19 April 2019

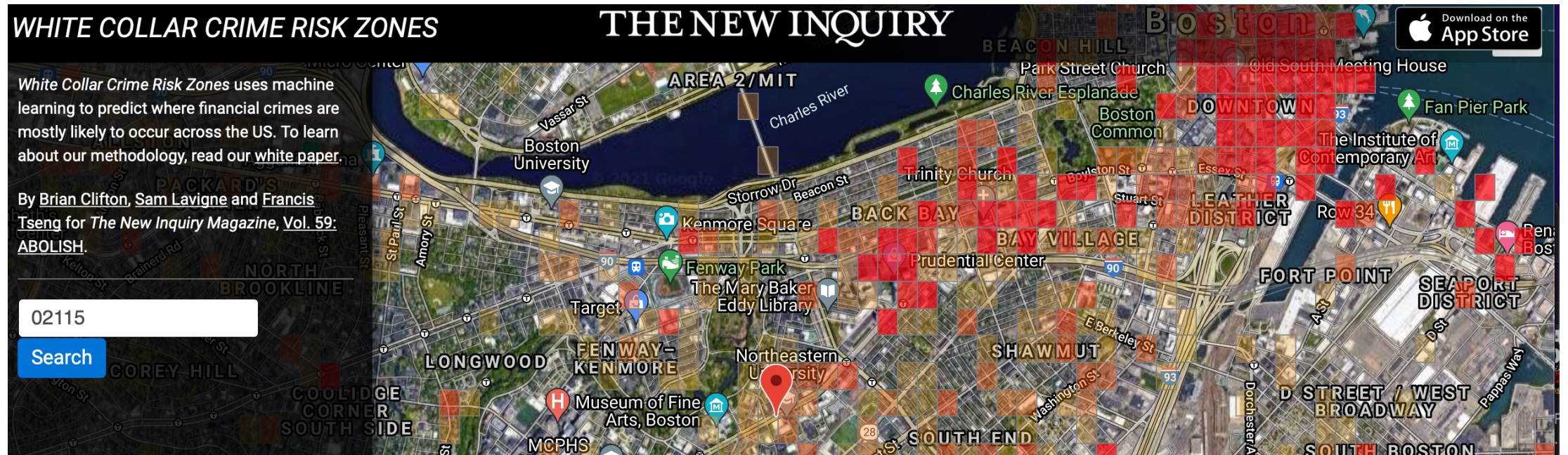
2. In Data Collection Practices (e.g., training data)

Sources of bad or biased training data

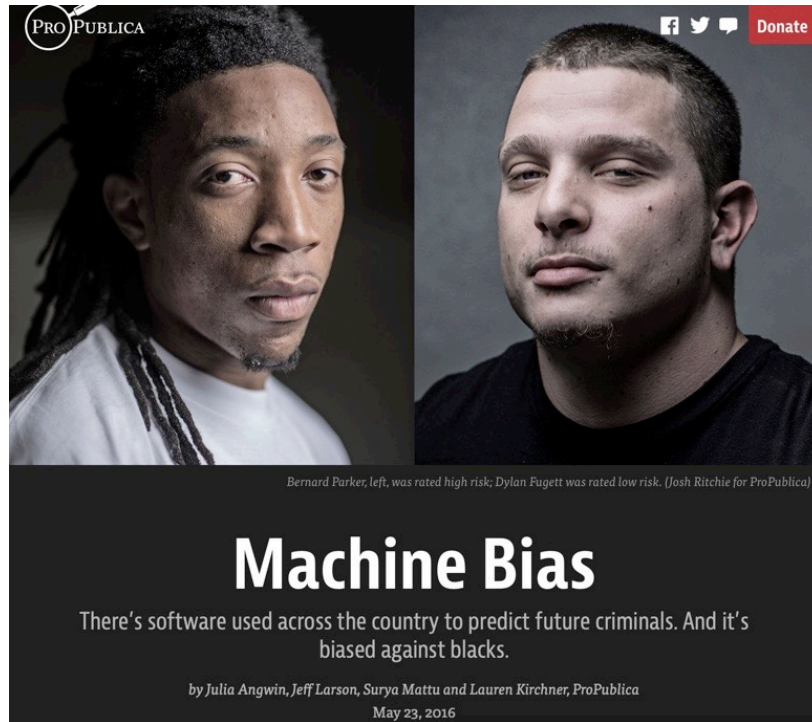
- a. When defining target variables and in class labels
- b. When assembling the training data set, resulting in an unrepresentative sample
- c. When selecting relevant features
- d. Intentional bias: masking, redlining, etc.



How are the categories defined?



3. In Distribution of Burdens of Algorithmic Error (in decisions or outcomes)



	White	Black
Labeled Higher Risk & Didn't Re-Offend (False +s)	23. %	44.9%
Labeled Lower Risk & Did Re-Offend (False -s)	47.7%	28.0%

Unfair distribution of error by racial class membership

ProPublica examined COMPAS risk assessment scores for about 7000 defendants in Broward County, FL. ProPublica found that while COMPAS correctly predicted recidivism 61% of the time, **the likelihood of different types of errors (FP and FNs) differed by race** (see table above)

Northpointe responded that COMPAS was nevertheless fair because *its positive predictions of recidivism were correct at the same rates regardless of racial group membership*; this response was followed by a detailed rebuttal by ProPublica

ML and Treating People as (Data) Subjects

The tension:

“constructing the human as a data point for machine training and optimization **rather than** as a person who should be justly, equitably, and sensitively treated”
(Chancellor et al., p 2)

What Ethics Is,
Why It
Matters,
and How It can
Help



What Ethics Isn't (Necessarily)

“It’s legal” ≠ “It’s ethical”



“It’s illegal” ≠ “It’s unethical”



What Ethics Isn't (Necessarily)





Ideals, aspirations, standards for how to live well and how to live well *together*



The uncovering and studying of those ideals and standards



The clarification, justification, and defense of those ideals and standards



The living by (or in accordance with) those ideals and standards

What Ethics Is

Are there answers to questions about how we ought to live, and what kinds of people we should strive to be?

Yes! But:

- There may not be *uniquely* correct answers to a given ethical question
- Even if there **are** uniquely right answers, it might be very difficult to find out what they are
- Details matter – e.g., historical context, social context, what values are relevant, and for whom

*A Pluralist and
Contextual
Approach To
Ethics*

Examples of Ethical Values - 1



Human welfare refers to people's physical, material, and psychological well-being



Accessibility refers to making all people successful users of information technology



Respect refers to treating people with politeness and consideration



Calmness refers to a peaceful and composed psychological state



Freedom from bias refers to systematic unfairness perpetrated on individuals or groups, including pre-existing social bias, technical bias, and emergent social bias

More Examples of Ethical Values - 2



Ownership and property refers to a right to possess an object (or information), use it, manage it, derive income from it, and bequeath it



Privacy refers to a claim, an entitlement, or a right of an individual to determine what information about himself or herself can be communicated to others



Trust refers to expectations that exist between people who can experience good will, extend good will toward others, feel vulnerable, and experience betrayal



Accountability refers to the properties that ensures that the actions of a person, people, or institution may be traced uniquely to the person, people, or institution

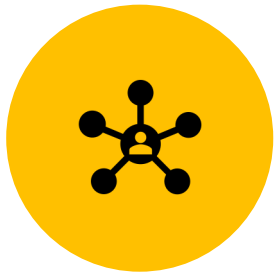
Even More Examples of Ethical Values - 3



Autonomy refers to people's ability to decide, plan, and act in ways that they believe will help them to achieve their goals



Informed consent refers to garnering people's agreement, encompassing criteria of disclosure, comprehension, voluntariness, competence, and agreement



Identity refers to people's understanding of who they are over time, embracing both continuity and discontinuity over time



Environmental sustainability refers to sustaining ecosystems such that they meet the needs of the present

Which 4 of these ethical values do you think are MOST important in an ML context, and why?

Accessibility

Accountability

Autonomy

Calm

Environmental
sustainability

Freedom from
bias

Human
welfare

Identity

Informed
consent

Ownership /
property

Privacy

Respect

Trust

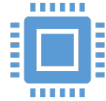


Introducing Value Sensitive Design (VSD)

The case for (the need for) VSD



Technology is
the result of
human
imagination



All technology
involves design



All design
involves
choices among
possible
options



All choices
reflects values



Therefore, all
technologies
reflect and
affect human
values



Ignoring values
in the design
process is
irresponsible

Three types of investigation in VSD

Empirical Investigation

- What do **stakeholders** say they value?
- How do stakeholders **interact with** this technology?
- What kinds of **financial/economic resources** does the technology require?

Value Investigation

- What is the **overall goal** of the technology?
- What **values** are at stake?
- Which stakeholders are **legitimately impacted**?
- What value-oriented criteria will be used to gauge project **success**?

Technical Investigation

- How can the tool or system be designed to enable designers to meet their value-oriented goals?
- What effect does **law, policy, and regulation** have on the design?
- Do the technical results **stay within your “red lines”**?

Value Sensitive Design (VSD) in action: the sequence

1. Who are the **stakeholders**? Identify them.

2. What **values** are at stake for those stakeholders? Identify them.

3. Where do there have to be “**tradeoffs**” between some values/interests and other values/interests?

4. Which **core values** need to be given priority, or “**red lines**” should not be crossed?


5. **Repeat** steps 1 – 4 as you get new information or as circumstances change.

Have a clear understanding of how the design can be **technologically successful**, not just technically successful.





1. Stakeholders: Whose values / interests are in question?

- **Direct** stakeholders include users, producers, and owners of the technology in question
 - **Indirect** stakeholders need to be assessed on a case-by-case basis (people who might not directly interact with the technology in question, but are affected by it nonetheless)
 - *Technologies affect more than just those who use them*
-



2. What are some of the values at stake in designing or using this technology?



3. What happens when values or interests come into conflict?

Value tradeoffs are needed when:

- multiple values are important;
- they also (seem) hard to achieve at the same time, and so
- a balance must be struck between them

Sometimes this might be different values held by the same party

- e.g., a company that values **security** but also **resource efficiency**
- e.g., should you be a programmer or a nurse?

Sometimes it might be the same value held by different parties

- e.g., **my financial interests** and **the tech company's financial interests**



4. (How) Can value tradeoffs be approached?

- **Assess legitimacy** → are everyone's interests equally legitimate in this context?
 - A burglar has a financial interest in your leaving your home unlocked...
 - **Respect core values and "red lines"** → are there any values that (almost) cannot be overridden?
 - For example: "No matter how much money they're offering you, you can't ..."
 - **Promote stronger values** → are there interests or "red lines" that should be prioritized in this context?
 - Increasing profit and preserving the environment might both matter, but preserving a habitable planet matters far more than increasing profit does
 - **Understand the social AND technical contexts** → Can some value tensions be revisited or resolved "later"?
-

6. What does a technologically successful solution look like?

“Success”: Technical v Technological

In CS, we typically think about **technical success**

- Does the technology function?
- Does it achieve first-order objectives?

Examples:

- Test coverage and bug tracker
- Crash reports
- Benchmarks of speed, prediction accuracy, etc.
- Counts of app installations, user clicks, pages viewed, interaction time, etc.

VSD asks that we think about **technological success**

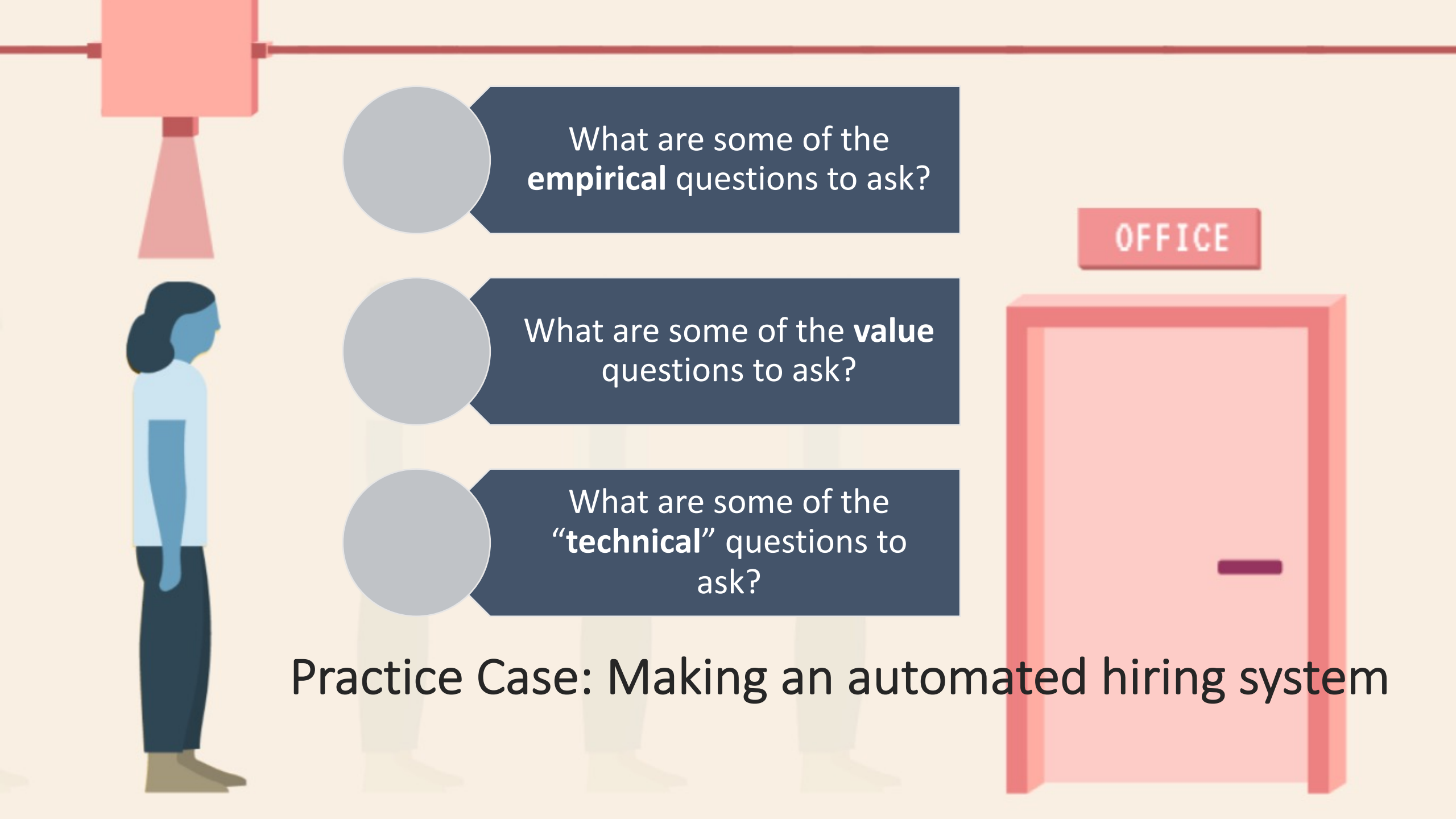
- Is the technology beneficial to stakeholders, society, the environment, etc.?
- Is the technology fair or just?

Examples:

- Assessments of quality of life
- Measures of bias
- Reports of bullying, hate speech, etc.
- Carbon footprint

Survival of the Best Fit, Revisited: Applying VSD

Remember: Value Sensitive Design is a heuristic framework to help you think about ethical issues, not an easy and unequivocal source of answers to ethical questions.

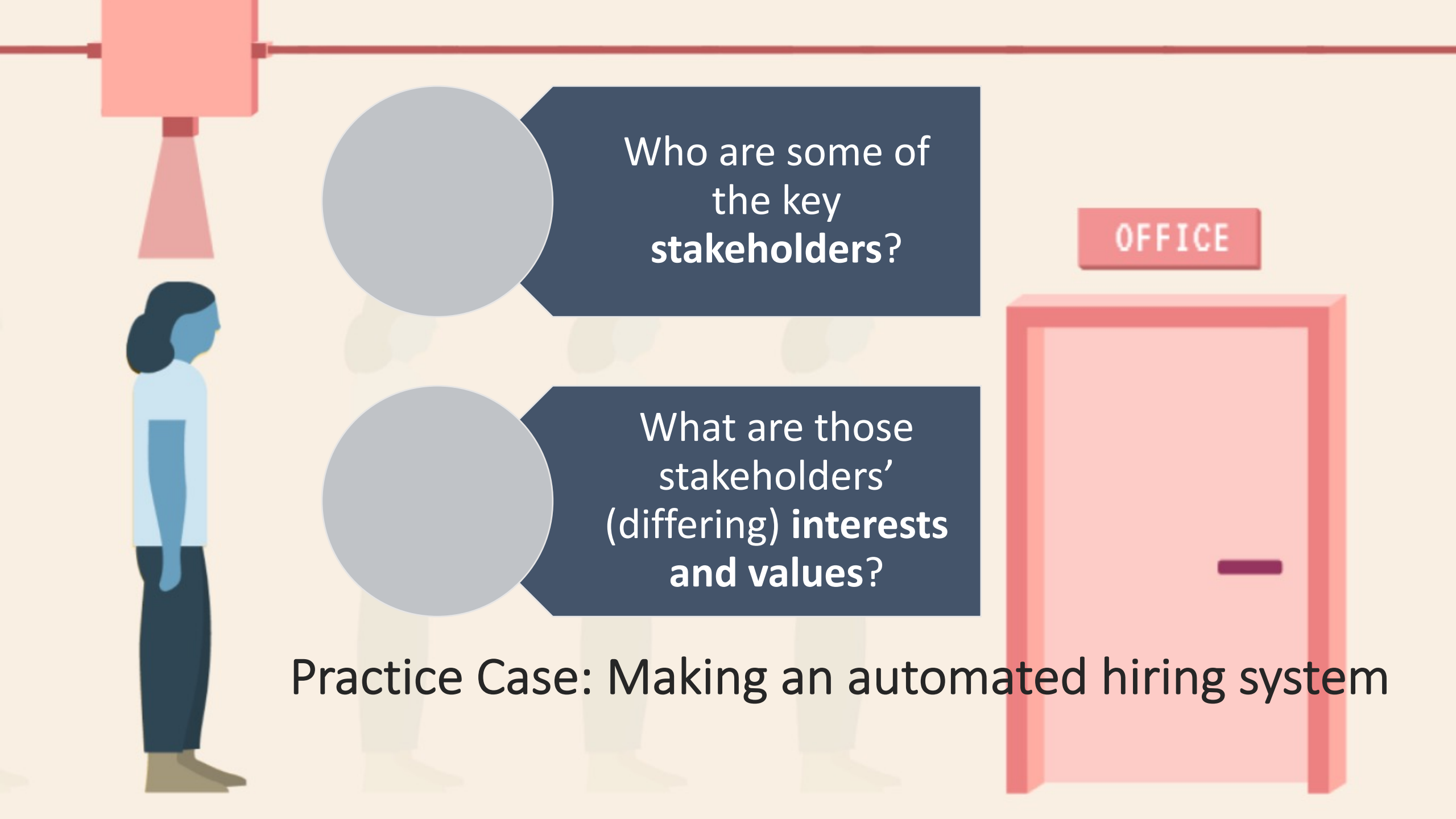
The illustration features a person with dark hair, wearing a light blue shirt and dark pants, standing on the left. Above them is a red rectangular object with a red cone pointing downwards. To the right of the person is a vertical list of three questions, each preceded by a grey circle. Further right is a red door with a small horizontal handle and a sign above it that says 'OFFICE'. At the bottom, the text 'Practice Case: Making an automated hiring system' is displayed.

What are some of the **empirical** questions to ask?

What are some of the **value** questions to ask?

What are some of the **“technical”** questions to ask?

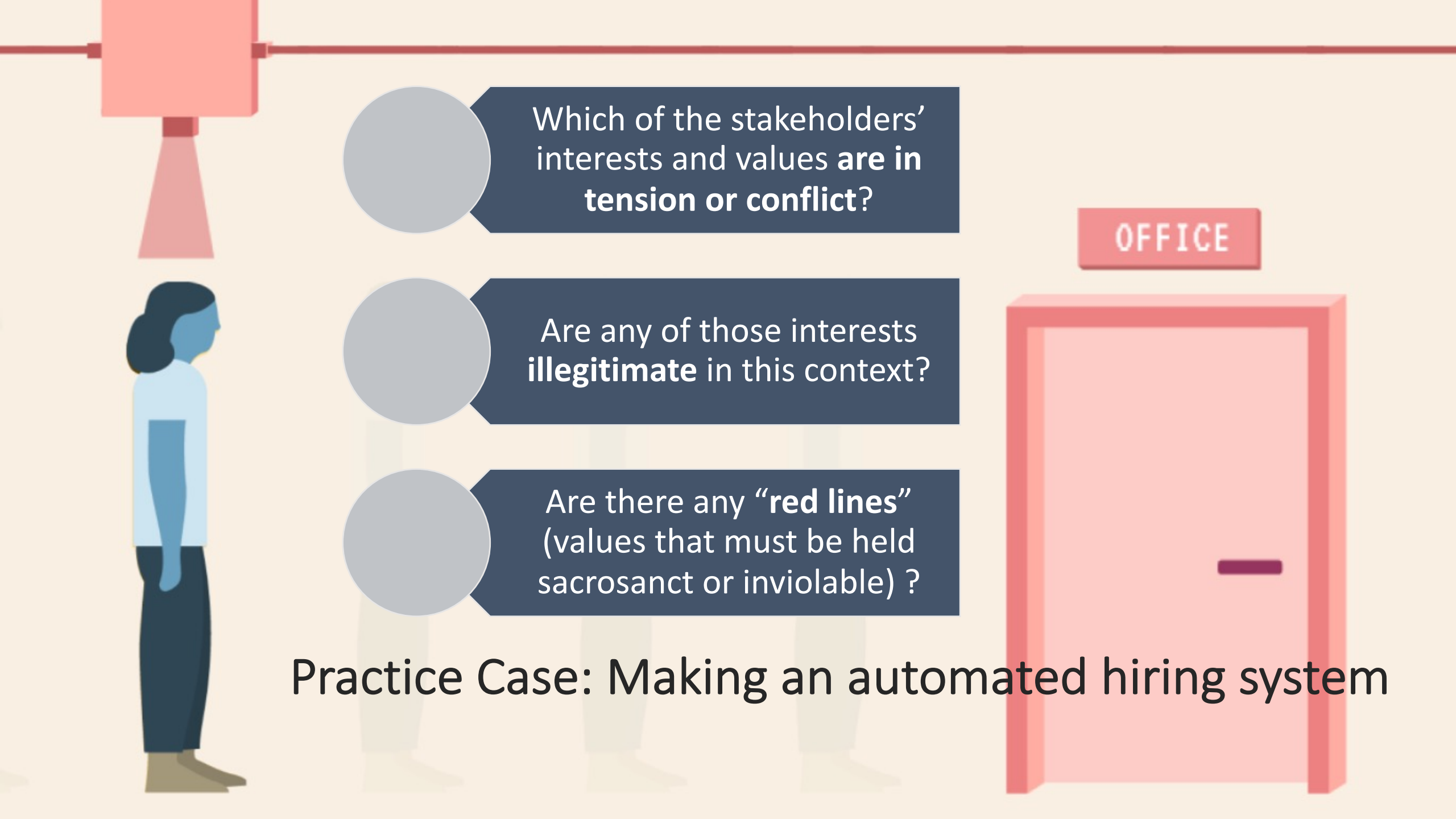
Practice Case: Making an automated hiring system

An illustration featuring a person in a light blue shirt and dark pants standing in a line of stylized figures. A red spotlight shines down on the person. To the right, a red door is labeled 'OFFICE' in white text. Two dark blue speech bubbles with white text are positioned in the center of the image.

Who are some of
the key
stakeholders?

What are those
stakeholders'
(differing) **interests**
and **values?**

Practice Case: Making an automated hiring system



Which of the stakeholders' interests and values **are in tension or conflict**?

Are any of those interests **illegitimate** in this context?

Are there any “**red lines**” (values that must be held sacrosanct or inviolable) ?

Practice Case: Making an automated hiring system



Value Sensitive Design (VSD) in action: the sequence

1. Who are the **stakeholders**? Identify them.
2. What **values** are at stake for those stakeholders? Identify them.
3. Where do there have to be “**tradeoffs**” between some values/interests and other values/interests?
4. Which **core values** need to be given priority, or “**red lines**” need to not be crossed?
5. **Repeat** steps 1 – 4 as you get new information or as circumstances change.
6. Have a clear understanding of a **successful outcome** of this process.

Determining “success” in this context

- 1) What is the objective of the KHASM algorithm?
- 2) What *features* and *data* might be useful to help train the algorithm to accomplish that objective?
- 3) How will you identify those features? How will you get that data?
- 4) How might those features and that data be illegitimately obtained; corrupted; unrepresentative; unjustifiedly biased?
- 5) How will you either prevent, or mitigate against, those problems?

Content Warning

brief mention of suicide and mental health crises



Quick Summary

- Crisis Text Line ([CTL](#)) is a prominent, not-for-profit service that “provides ‘mental health crisis’ intervention services” to people seeking help
- Before hotline users seeking assistance speak to volunteer counselors, they consent to data collection (and they can read the company’s data-sharing practices)
- CTL says it uses that data to “help identify the neediest cases or zero in on people’s troubles, in much the same way that Amazon, Facebook and Google mine trends from likes and searches”

Quick Summary, continued

- CTL entered into an agreement with its for-profit spinoff company, Loris.ai
- According to the agreement, CTL would provide Loris.ai “sliced, repackaged, and anonymized” data from its online text conversations with its clients
- Loris.ai would use that data in a training set to help design a ML-based algorithm for customer service software that it would sell to other companies
- CTL would receive a share of the profits from those sales

More practice
with VSD:
the case of Crisis
Text Line (CTL)
and Loris.ai

TECHNOLOGY

Suicide hotline shares data with for-profit spinoff, raising ethical questions

The Crisis Text Line's AI-driven chat service has gathered troves of data from its conversations with people suffering life's toughest situations.



Politico, January 28, 2022

What was the situation?

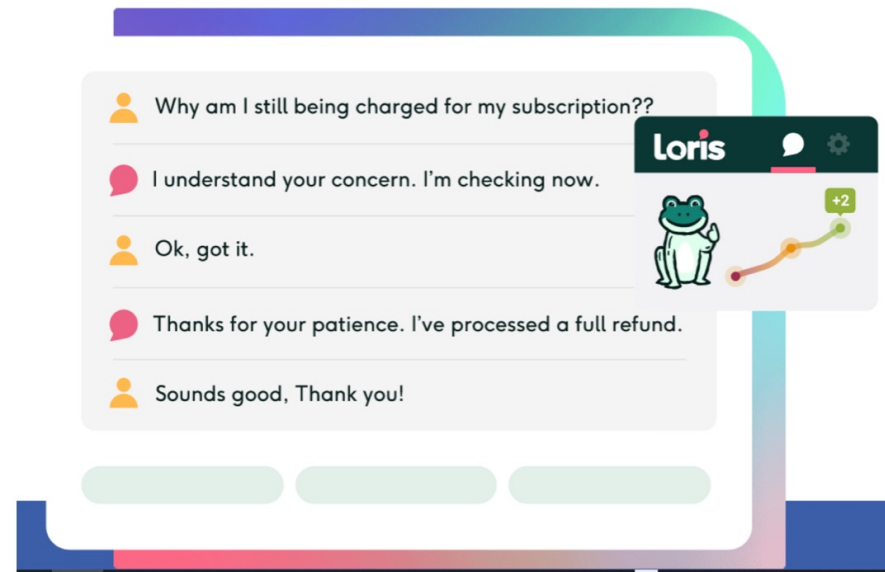
Crisis Text Line is one of the world's most prominent mental health support lines, a tech-driven nonprofit that uses big data and artificial intelligence to help people cope with traumas such as self-harm, emotional abuse and thoughts of suicide.

But the data the charity collects from its online text conversations with people in their darkest moments does not end there: The organization's for-profit spinoff uses a sliced and repackaged version of that information to create and market customer service software.

Crisis Text Line says any data it shares with that company, [Loris.ai](#), has been wholly “anonymized,” stripped of any details that could be used to identify people who contacted the helpline in distress. Both entities say their goal is to improve the world — in Loris' case, by making “[customer support more human, empathetic, and scalable](#).”

What's was the situation (continued)?

In turn, Loris has pledged to share some of its revenue with Crisis Text Line. The nonprofit also holds an ownership stake in the company, and the two entities shared the same CEO for at least a year and a half. The two call their relationship a model for how commercial enterprises can help charitable endeavors thrive.



The website of the company Loris.ai offers this example of how its software uses artificial intelligence to help customer service agents interact with consumers online. | Loris' website

Where do things stand now?

Crisis Text Line has decided to stop sharing conversation data with spun-off AI company Loris.ai after facing scrutiny from data privacy experts. “During these past days, we have listened closely to our community’s concerns,” the 24/7 hotline service writes in a [statement](#) on its website. “We hear you. Crisis Text Line has had an open and public relationship with Loris AI. We understand that you don’t want Crisis Text Line to share any data with Loris, even though the data is handled securely, anonymized and scrubbed of personally identifiable information.” Loris.ai will delete any data it has received from Crisis Text Line.

The Verge, February 01, 2022

The question of secondary uses and meaningful consent

1. Do you know where the data came from in your training set?
2. Can any of us be informed about the possibility that “our data” will end up in someone else’s training set?
3. What ethical principles should be followed when assembling a dataset? How can people whose data is included, participate meaningfully in that process?
4. Are they entitled to expect to be able to participate meaningfully?

Value Sensitive Design (VSD) in action: the case of CTL

1. Who are the **stakeholders**? Identify them.

2. What **values** are at stake for those stakeholders? Identify them.

3. Where do there have to be “**tradeoffs**” between some values/interests and other values/interests?

4. Which **core values** need to be given priority, or “**red lines**” should not be crossed?

5. **Repeat** steps 1 – 4 as you get new information or as circumstances change.

Have a clear understanding of how the design can be **technologically successful**, not just technically successful.

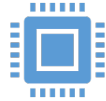
Review and conclusion

A vertical line is positioned to the right of the text. In the bottom right corner of the slide, there is a large yellow triangle pointing upwards and to the left, partially overlapping a light gray rectangular area.

The case for VSD, One More Time



Technology is
the result of
human
imagination



All technology
involves design



All design
involves
choices among
possible
options



All choices
reflects values



Therefore, all
technologies
reflect and
affect human
values



Ignoring values
in the design
process is
irresponsible


How do we avoid (creating or relying on machine learning algorithms that end up) treating people unfairly?

First, *explain why this system should be built in the first place*

- What makes you think that the problem (or set of problems) *can* be satisfactorily addressed by software, especially ML?
- What makes you think that the problem (or set of problems) *should* be satisfactorily addressed by software, especially ML?
- What are the social, political, and moral contexts in which this system will be used?
- What will happen when a user you didn't imagine, uses your system in a way that you didn't expect?

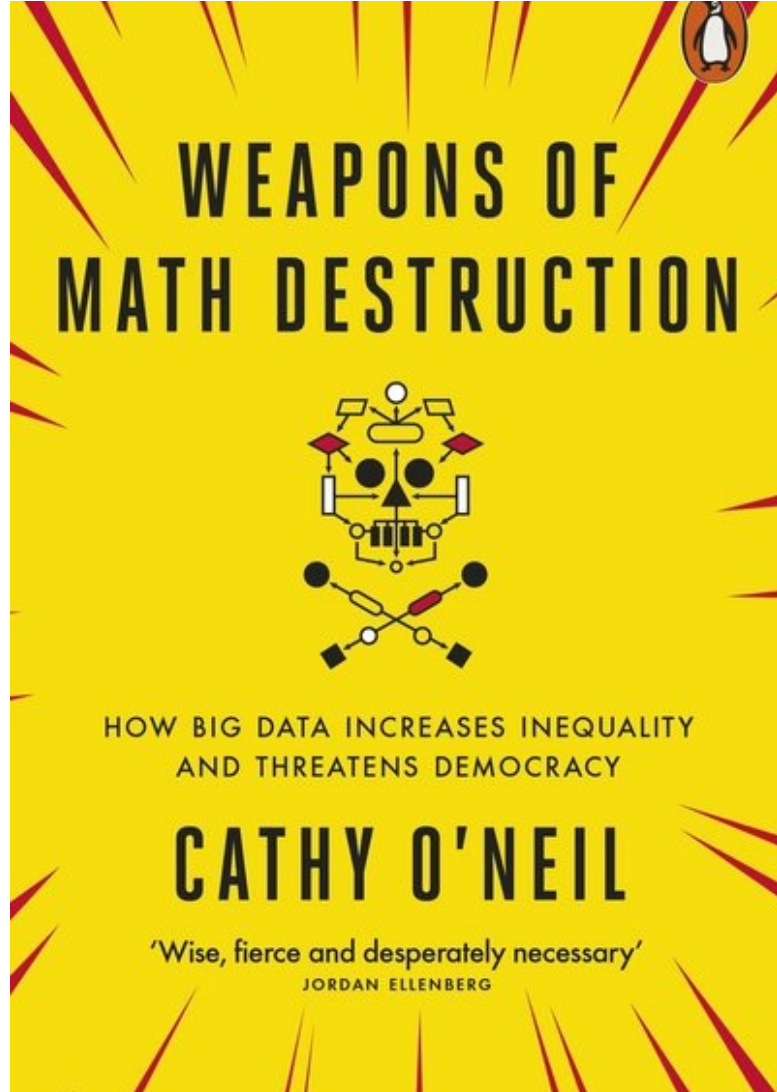
*Second, pay
careful
attention to
how training
data is
collected*

- When defining target variables and in class labels
- When assembling the training data set, resulting in an unrepresentative sample
- When selecting relevant features
- Watch out for **intentional** bias: masking, redlining, etc.



Third, make
explicit ethical
decisions about
how to distribute
the risks/results
of algorithmic
error

- Even if the algorithm is perfectly accurate (?), its results or use might violate important ideals (e.g., of fair treatment)
- To distribute the risks of error ethically, you should bring in all stakeholders (designers, users, “bystanders”, etc.) for collaboration about definition/selection of relevant features, refinement of the training data set, etc.
- Again: consider whether a machine learning algorithm *should be used at all* in this domain



We must, therefore, make careful, explicit choices as to how and where to distribute the burdens of error in the algorithms we build.

This should be done at both the **law and policy** level, and at the **design** level, which is where value-sensitive design – an approach that emphasizes stakeholder interests and values – attempts to intervene.

We should also ask *whether an algorithm should be used at all* for the task at hand.

VSD: An outlook, not a (mindless) algorithm

Fundamentally, VSD is an **outlook** and a **process**

- VSD is not an algorithm
- There is no design recipe for VSD
- There is no way to *#include vsd.h* or *import VSD*

Committing to VSD means being **thoughtful** and **responsive**

- No single right answer to complex ethical and moral questions...
- But there are **lots of wrong answers**

Engaging with values in the design process offers creative opportunities for:

- Technical innovation
- Improving the human condition (*doing good and saving the world*)