Learning from the People

From Normative to Descriptive Solutions to Problems in Security, Privacy & Machine Learning

Elissa M. Redmiles, Microsoft Research & Max Planck Institute for Software Systems

@eredmiles
eredmiles@gmail.com
Computational problems require constant decision-making.
Typically: experts set best practices
Experts trade off costs and benefits

Benefit to science

Risk to people
Unfairness
Burden

Realm of possibility

Expert’s Normative Decision

security
classifier accuracy
Experts do not always agree on best practices.

A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

The future of artificial intelligence: two experts disagree.

Security

Widely Used Password Advice Turns Out to Be Wrong, NIST Says

New recommendations from the National Institute of Standards and Technology call for people to create passwords that are 'easy, easy to remember phrases' -- a series of four to five words mixed together.

By Douglas Ferguson, The Oregonian, Portland, OR, Aug 9, 1017
More importantly, users and experts may disagree.
This disagreement is a classic tension in moral philosophy

<table>
<thead>
<tr>
<th>Normative</th>
<th>Descriptive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experts <strong>prescribe</strong> best practices</td>
<td><strong>Learn</strong> non-expert preference/behavior</td>
</tr>
<tr>
<td></td>
<td><strong>Infer</strong> best practices</td>
</tr>
</tbody>
</table>
Can we use descriptive approaches in computational decision-making?
Three case studies, three different descriptive methods

1. How should we set **security** policies?
   - Observe behavior
   - Infer preference
   - Make decision

2. Which features are fair to use in **machine learning**?
   - Ask preference
   - Make decision

3. What content should be allowed in virtual reality?
   - Make decision **together**
Security

Determine how & when to prompt secure behavior

Goal
Get users to behave more securely by prompting

Protect your account with 2-Step Verification

Each time you sign in to your Google Account, you'll need your password and a verification code.

Learn more

Add an extra layer of security
Enter your password and a unique verification code that's sent to your phone.

Keep the bad guys out
Even if someone else gets your password, it won't be enough to sign in to your account.

Google 2-step verification
Image credit: EFF 2016
Why don’t users behave as expected when prompted?

The user is going to pick **dancing pigs** over **security** every time.

-- McGraw and Felten / Schneier
Measure prompt response using a novel, scalable behavioral-economics security experimentation system

Online experimental system: simple bank account
Account holds study compensation
Account has explicit risk of being hacked

Participants interact with simulation system
We observe their responses to security prompts

At the end of the study, you will be compensated with the amount of money left in your study bank account. **You begin the study with $1 each day that you login you will earn an additional $1, up to a total of $5.** You must login once a day, otherwise you will lose all of the money in your account. If you are hacked, you will also lose all of the money in your account.

Studies indicate that 20% of users will have their study accounts hacked over the course of the year.

Would you like to enable two factor authentication using your phone number? Two factor authentication will protect you from hacking 90% of the time.

**H** = \( N\% \)

**P** = \( N\% \)

You will lose all of your money if you do not login before January 19, 2018, 5:02pm EST.

Bank: $5
Only 52% of participants enabled 2FA.
Testing the bounded rationality hypothesis: is there a consistent pattern in security behavior?

Enable 2FA  ~  Account value  +  Risk with/out 2FA

Controls
- Password Strength
- Internet & Security Skill
- Demographics (Gender, Age, Education)

Neural Net
- Strength Meter
- Scales

Validated
- Hargittai & Hsieh 2013
- Egelman & Peer 2015

Testing the bounded rationality hypothesis: is there a consistent pattern in security behavior?

Enable 2FA

Costs proxy: time spent

Past Behavior (RD1 2FA choice)

Experimental results suggest users are boundedly rational.

\[ \text{Risk (H, P) + Account Value (Earn/Endow)} \]

e explains 9% behavior variance.

People behave in ways we can model well

We can model human behavior well ($R^2=0.61$) as a function of variables measured or controlled in the simulation system.

Differences in ability (differences in cost) alter behavior

Differences in account valuation alter behavior

Normative
Prompt everyone to use 2FA until they do: it’s good for them
Problem: people are so inundated they start ignoring prompts
Problem: not everyone gets the same value out of the same behavior

Inequalities in Ability (e.g., 2FA difficulty)
Valuation of account

Allocate Resources

Customize Messages
Can we use our descriptive knowledge to set prompts?

How should we set **security** policies?
- Observe behavior
- Infer preference
- Make decision

Which features are fair to use in **machine learning**?
- Ask preference
- Make decision

What content should be allowed in virtual reality?
- Make decision **together**
Mechanism design to facilitate descriptive approach
Companies can maximize profit by selecting optimal values for factors they control.

$$\max \text{profit} \quad B_s, B_q, m, r$$

s.t.

$$C - \text{budget} \leq 0 \quad [B_s, B_q] < \epsilon$$

where

$$\text{profit} = \sum_{i=1}^{n} [g(u_i) - c(b_i, r_i)] - c(B_s, B_q).$$
Mechanism design enables descriptive approach and introduction of equity notions.

Inequalities in Ability (e.g., 2FA difficulty)

Valuation of account

Allocate Resources

\[ m_1 \text{ for } u_i \]

\[ r_1 \text{ for } u_i \]

Constraints

- Inequalities in Ability (e.g., 2FA difficulty)
- Effort equity: minimize variance in costs
- Risk equity: minimize variance in risk

Security Goal

Customize Messages

Allocate Resources

Linear Programming

\[
\begin{align*}
\max & \quad g(u) \\
\text{s.t.} & \quad C - \text{budget} \leq 0 \\
& \quad [B_s, B_q] < \epsilon \\
& \quad \text{where } \text{protection} = \sum_{i=1}^{n} [q(u_i) - c(b_s, b_q)] - c(B_s, B_q).
\end{align*}
\]
Decide by solving an optimization problem that uses knowledge of user behavior gained through observation.

- **How should we set security policies?**
  - Observe behavior
  - Infer preference
  - Make decision

- **Which features are fair to use in machine learning?**
  - Ask preference
  - Make decision

- **What content should be allowed in virtual reality?**
  - Make decision together
Can we get to a decision sooner? Directly ask the users

- How should we set **security** policies?
  - Observe behavior
  - Infer preference
  - Make decision

- Which features are fair to use in **machine learning**?
  - Ask preference
  - Make decision

- What content should be allowed in virtual reality?
  - Make decision **together**
Goal
Determine which features are fair to use in a classifier

Descriptive Approach
Model how users reason about fairness and include/weight features based on fairness judgements

Machine Learning
Select features that are fair to use for classification
Let’s back up for a moment: why do we care about feature fairness?

Google’s algorithm shows prestigious job ads to men, but not to women. Here’s why that should worry you.

By Julie Carpenter July 6, 2015

Google’s image searches for “CEO” only turn up pictures of white men, there’s new evidence that algorithmic bias is, alas, at it again. In a paper published in April, a team of researchers from Carnegie Mellon University claim Google displays far fewer ads for high-paying executive jobs...
Systemy is a local technology firm that develops software. They are expanding and want to hire new employees. Systemy contracts with Bezo Media, an online advertising network, which places Systemy’s job ad on a local news website. An HR employee at Systemy chooses to target individuals who are Asian rather than individuals who are White. Asian individuals tend to click on different ads than individuals who are White. Bezo Media’s automated system has observed this difference and automatically assigns the Systemy ads to Asian individuals rather than individuals who are White. As a result, the ad is shown more frequently to individuals who are Asian than who are White.

What drives perceptions of ad discrimination scenario?

Measured the effect of varying beneficiary, targeting mechanism & targeted features

Training Data Collection
MTurk survey (n=191) for training regression models

Final Survey & Modeling
Census-representative web panel sample (n=891) with 5-fold CV on trained models
Features are a key factor of perceived fairness

Fairness perception is based on the features (demographic vs. behavior)

- **Explicit Demographic**
- **Behavior Inference**

<table>
<thead>
<tr>
<th>Mechanism of Targeting</th>
<th>Majority Group Benefit</th>
<th>Minority Group Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>Not at all a problem</td>
<td>Minor problem</td>
</tr>
<tr>
<td>Algorithm</td>
<td>Not at all a problem</td>
<td>Minor problem</td>
</tr>
<tr>
<td>Human</td>
<td>Moderate problem</td>
<td>Serious problem</td>
</tr>
<tr>
<td>Algorithm</td>
<td>Moderate problem</td>
<td>Serious problem</td>
</tr>
</tbody>
</table>

COMPAS system helps Florida judges make bail decisions

Output
Chance of recidivism
- HIGH
- MED
- LOW

Recidivism Risk
Predict recidivism risk from questionnaire answers

**Input**
Defendant’s answers to COMPAS questionnaire

**Features**
Selected answers to questions

- Current charge
- Criminal history of family and friends
- Performance in School
- Mental health status
- Nothing Legally Sensitive (Race, Gender, etc.)

**Output**
Chance of recidivism

- HIGH
- MED
- LOW

Unfair Features
Analog system: judges admit evidence

Features
Selected answers to questions

Unfair Features
COMPAS: algorithm designers select features

**Features**
Selected answers to questions

Algorithm Designer

Unfair Features
What If?
We Followed People’s Beliefs About Fairness

Features
Selected answers to questions

Unfair Features
Survey to assess people’s fairness beliefs

Online survey

Judges in Broward County, Florida, have started using a computer program to help them decide which defendants can be released on bail before trial. The computer program they are using takes into account information about the defendant’s stability of employment and living situation.

For example, the computer program will take into account the defendant’s answer to the following question: **How often do you have trouble paying bills?**

Please rate how much you agree with the following statement:
It is fair to determine if a person can be released on bail using information about their stability of employment and living situation.
Human Perceptions of Fairness in Algorithmic Decision Making.

The Web Conference (WWW2018).
Lack of consensus on fairness beliefs, why?

- **How should we set security policies?**
  - Observe behavior
  - Infer preference
  - Make decision

- **Which features are fair to use in machine learning?**
  - Ask preference
  - Make decision

- **What content should be allowed in virtual reality?**
  - Make decision **together**
People determine ``fairness” based on eight sub-questions:

- Reliable?
- Relevant?
- Private?
- Volitional?
- Causes Outcome?
- Causes Vicious Cycle?
- Causes Disparity in Outcomes?
- Caused by Sensitive Group Membership?

Legal:
- admissible evidence

Philosophical:
- Causal Reasoning

Sociological:
- disparate impact

Political Science & Economics

88% accuracy predicting fairness from property ratings

Lack of consensus in property ratings, not fairness beliefs

Descriptive for mapping properties to fairness
Normative to evaluate feature properties

Far in the Future
Computationally evaluate properties

Normative
Judges/experts evaluate properties

Reliable: 6
Relevant: 2
Private: 5

Reliable: 6
Relevant: 2

Reliable: 3
Relevant: 5
Private: 5
Volitional: 2

Descriptive Mapping Function
From Properties to Fairness

How should we set **security** policies?

- **Observe** behavior
- **Infer** preference
- **Make decision**

Which features are fair to use in **machine learning**?

- **Ask** preference
- **Make decision**

What content should be allowed in virtual reality?

- **Make decision together**

Constrain search for features based on fairness threshold

Can we just make the decision together with the users?

- **How should we set security policies?**
  - Observe behavior
  - Infer preference
  - Make decision

- **Which features are fair to use in machine learning?**
  - Ask preference
  - Make decision

- **What content should be allowed in virtual reality?**
  - Make decision together
Virtual Reality

Define guidelines for VR development

Goal
Determine guidelines for developing VR content

Descriptive Approach
Co-design standards with VR developers
Interview Study: VR developers want guidelines

“there’s a quite a big list of unknowns right now in terms of what’s best etiquette for a user and what’s gonna keep the user the most [safe], comfortable, and satisfied”

-- Developer 8

“just the fact of the matter is there are no VR power users. I can count on my hand the number of experienced ‘devs’ I’ve actually met”

-- Developer 5

Code of ethics co-design with developers

Six high level principles drawn by researchers from interview results

Invite 11 online communities of VR developers to edit the draft
Standards for Ethical Development in VR

Do No Harm. We will ensure that the intensity of VR experiences is appropriate by thorough testing.

Secure the Experience. We will use the best security protocols and protections of which we are aware to ensure that malicious actors cannot alter or harm a users’ experience while they are in VR.

Be Transparent About Data Collection. We will ensure that our privacy policies specifically mention VR data and how that data will be used (and shared) and protected.

Ask for Permission. We will include permission requests, if at all possible, for sensitive data such as eye-tracking information, health or biometrical information, including movement-derived data.

Keep the Nausea Away. We will test all products before release and do our best to reduce nausea among our users.

Diversity of Representation. We will work to ensure that a diverse array of avatars are available for use by users and that our representations of groups and characters does not perpetuate stereotypes.

Social Spaces. We will take extra care through privacy protections and clear and conscious community guidelines and moderation affordances to ensure that cyberbullying and sexual harassment is kept to a minimum and social VR experiences are kept safe and inclusive.

Projects involving children (or other vulnerable populations) deserve special consideration.

Accessibility for All. Include options for those without standard vision, hearing, or movement to enable them to participate meaningfully in experiences, for example through modular design that allows users to integrate additional software or hardware as needed. As long as it doesn’t hurt the vision of the project, the idea of the project comes first.

User-Centric User Design and Experience. Make good UX that is designed to be informative to end users.

Proactive Innovation: We will seek out and implement relevant methods by which to enhance, immerse and make seamless the experience in which we provide for our users. This includes the acknowledgement that we as an entity are inclusive of our ecosystem and not separate from it in relation to our end-users and act as a unifying body in collaboration and symbiosis for the best possible experience overall.

Developers reached consensus on 10 principles

Six high level principles drawn by researchers from interview results

Invite 11 online communities of VR developers to edit the draft

Trace ethnographic analysis of editing process [see paper]
Standards for Ethical Development in VR

**Do No Harm.** We will ensure that the intensity of VR experiences, and effects caused (e.g., seizure risk from flashing lights) is appropriate by thorough testing. Avoid creating content that objectifies, demeans or violates the rights of humans or animals (e.g., creating experiences considered illegal or morally reprehensible if experienced in “real life”).

**Secure the Experience.** We will use the best security protocols and protections of which we are aware to ensure that malicious actors cannot alter or harm a users’ experience while they are in VR.

**Be Transparent About Data Collection.** We will ensure that our privacy policies specifically mention VR data and how that data will be used (and shared) and protected.

**Ask for Permission.** We will include permission requests, if at all possible, for sensitive data such as eye-tracking information, health or biometrical information, including movement-derived data.

**Keep the Nausea Away.** We will test all products before release and do our best to reduce nausea among our users.

**Diversity of Representation.** We will work to ensure that a diverse array of avatars are available for use by users and that our representations of groups and characters does not perpetuate stereotypes.

**Social Spaces.** We will take extra care through privacy protections and clear and conspicuous community guidelines to ensure that cyberbullying and sexual harassment is kept to a minimum and social VR experiences are kept safe and inclusive. Projects involving children or other vulnerable populations deserve special consideration.

**Accessibility for All: Include options for those without standard vision, hearing, or movement to enable them to participate meaningfully in experiences, for example through modular design that allows users to integrate additional software or hardware as needed.**

**User-Centric User Design and Experience.** Make good UX that is designed to be informative to end-users.

**Proactive Innovation:** We will seek out and implement new methods to enhance the immersive and seamless experience we provide to our users. We will not consider end-users as entirely separate; we will get in collaboration and ambience with them to achieve the best possible experience.
Different methods are appropriate for different problems
Methods have prerequisites: observation and question-asking require consistency.
Co-design requires recruiting users you think will make “good” choices or A LOT of users
Why Not Have VR Users Co-Design, Too?
Researchers normatively decided that small group of users with homogenous, exclusive opinions weren’t good first-round participants

...If you use VR, most likely you [also use] Reddit because there’s a certain type of crowd that’s really into this, you know?

“somebody who has a lot of money and has a premium setup you know...I mean you are talking people with 4 plus sensors.

I’ll be more concerned about virtual crimes and bullying once VR becomes more accessible to the “general public.”

Users
Descriptive vs. Normative: always a balance

Security
Normative expert effectiveness judgement
Future: compute effectiveness

Machine Learning
Normative expert property judgements
Future: compute property values

Virtual Reality
Normative researcher judgement of who to include in descriptive approach

At what are the humans best?
the experts
the computing systems
Explore descriptive solutions to computational problems: learning best practices from people’s preferences / behavior.

Through examples in security, machine learning, and virtual reality.

Illustrate how different balances between normative & descriptive could be achieved.

Learning from the People
From Normative to Descriptive Solutions to Problems in Security, Privacy & Machine Learning

Elissa M. Redmiles
@eredmil1
eredmiles@gmail.com