Privacy, Triage, and Risk
Or, The Art of VP-Wrangling

Yonatan Zunger
Distinguished Engineer & Chief Ethics Officer, Humu
zunger@humu.com

PEPR 2019
My Key Collaborators

- Lea Kissner – Privacy and respect in computing
- A. V. Flox – Psychological safety, activation, and trauma
- The Google Privacy Team – years of work and collaboration
The Talk I’m Not Giving: “Building for Trust”

- What is building for trust?
- Why build for trust?
- How do we build for trust?
- How do we assure ourselves we’ve build correctly?
- Where are the big challenges?

For the broad outline, see Lea Kissner’s IWPE 2019 Keynote, or Sha Sundaram’s PEPR talk from yesterday.
The Talk I’m Not Giving: “Building for Trust”

- What is building for trust?
- Why build for trust?
- How do we build for trust?
- How do we assure ourselves we’ve build correctly?
- Where are the big challenges?

For the broad outline, see Lea Kissner’s IWPE 2019 Keynote, or Sha Sundaram’s PEPR talk from yesterday.
How do we build for trust?

1. Find all the parts.
2. Check the invariants.
3. Model the threats.
4. Triage and correct issues.
How do we build for trust?

1. Find all the parts.
2. Check the invariants.
3. Model the threats.
4. Triage and correct issues.
Triage includes three things

1. Identify possible responses to threats
2. Prioritize threat response (vs all other tasks)
3. Achieve consensus on prioritization
Triage includes three things

1. Identify possible responses to threats
2. Prioritize threat management (vs all other tasks)
3. Achieve consensus on prioritization

Requires common language, alignment of understanding

Good tools for (3) help (2), and vice-versa.
Why triage is hard

1. Things people estimate poorly:
   
   Rare, catastrophic events.

   Continuous invisible impacts that add up.

2. People are drowning in priorities.

3. Different people have different threat models.
Four useful tools

1. The Universal Value Curve
2. The Gottman Ratio
3. Sense vs. Presence
4. Activists vs. Operators
The Universal Value Curve

How do you trade product value vs. rare catastrophic risks?

![Value Curve Diagram](image)
The Universal Value Curve

Applied to a single user of a specific product

- Your product doesn't exist
- Your product makes people's lives better
- Even the best tool has finite value to any person
- But things can go arbitrarily badly wrong
You’ve seen this curve before.

“Money can’t buy you happiness, but poverty can buy you one hell of a lot of misery.”
The Universal Value Curve

This also helps understand why this curve works

Money buys safety
Mitigate risks

Emotional homeostasis
Diminishing marginal value
Lottery Effect


See also: Notorious B.I.G., “Mo’ Money, Mo’ Problems,” Life After Death
Most happiness comes from eliminating stressors

See also: Notorious B.I.G., “Mo’ Money, Mo’ Problems,” Life After Death
The Universal Value Curve

Applied to the total user population

1. I just watched someone get harassed on my product
2. I just heard from a user whose job I saved!
3. I just made someone infinitely happy!
4. I just read the coroner’s report
That may seem obvious, but...

Do your product success metrics increase unboundedly per user?

Metrics like “total engagement-seconds” can create **perverse incentives**. Was the user any happier?

Do the numbers that drive your company align with user happiness? In the short term? In the long term?
The Universal Value Curve

Use this to **assess risk, design metrics, and convey tradeoffs**

Far LHS: Don’t multiply.  
bignum × smallnum = noise  
You can’t measure accurately enough.

**Instead: Count, use narrative**  
Graph is *qualitative, illustrative*  
“How many deaths per year is this feature worth?”

RHS: Be quantitative.  
Examine unbounded metrics *carefully*.  
Compare actively alienated users to users acquired.  
Talk PR risk, legal risk, regulatory risk, human harm.
“Relationships need to have at least a 5 to 1 ratio of positivity to negativity during conflict”

Applied to user experience:

● Positive moments (“aha!”, task accomplished, pleasure) versus

● Negative moments (stuck, broken, harassment)

The Gottman Ratio

What’s actually happening?

- The body is designed to detect danger cues and go on alert (neuroception).
- “Negative moments” are unsafety signals: “did I just lose my file?” “Am I being attacked?”
  - Same kind of unsafety as in LHS of value curve!
  - Response via autonomous NS, 50-500msec; decay via meta-cognitive processes, seconds-hours.

The Gottman Ratio

- As emotional resources deplete, the trigger threshold drops
  - People stop looking for safety and start confirming danger.
- Creates a **positive-feedback loop**:
  - Initial susceptibility $\propto$ person’s baseline unsafety.
- A single source can trigger the signal many times
  - Re-reading alarming text
  - Encountering other people’s danger responses
  - Attempt to remedy which fails

A. V. Flox, *Disrupting the Bystander*, Thorntree Press, 2019, ch. 3
Consequence: When people feel unsafe, they’re...

- Less likely to explore, learn, exchange information
- Less likely to keep using the product
  - 30% of people who have experienced serious online harassment have stopped using an online service after witnessing harassment there.
  - 18% have experienced serious harassment
  - 41% of 18-to-29-year-olds

Y. Gong et al., “Unfolding the Proactive Process for Creativity,” J. Mgt 2012 38: 1611

The Gottman Ratio

Use this to **design interfaces, highlight user loss risk.**

- Measure positive/negative ratio directly, when you can.
  - This can be a good user happiness metric
  - Gottman ratio hints at where it saturates
- When something goes wrong, minimize time until user can act to relieve the problem.
  - Easy report flow; quick help; get back to “safe state.”
  - Think 100s of msec, not minutes.
## Sense vs. Presence

How do you prioritize user worries versus real dangers?

<table>
<thead>
<tr>
<th>Sense of Threat</th>
<th>Presence of Threat</th>
<th>Safe</th>
<th>Alert</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No Threat</strong></td>
<td>Safe</td>
<td>Everything is great! <em>Relax</em></td>
<td></td>
</tr>
<tr>
<td><strong>General Danger</strong></td>
<td>Safe</td>
<td>People afraid and limiting behavior <em>Communicate</em></td>
<td></td>
</tr>
<tr>
<td><strong>Invisible Danger</strong></td>
<td>Presence of Threat</td>
<td>People unaware of threat <em>Mitigate danger Communicate with care</em></td>
<td></td>
</tr>
<tr>
<td><strong>Clear Danger</strong></td>
<td>Bad</td>
<td>People aware of real danger <em>Mitigate and communicate</em></td>
<td></td>
</tr>
</tbody>
</table>
### Sense vs. Presence

How do you prioritize user worries versus real dangers?

<table>
<thead>
<tr>
<th>Presence of Threat</th>
<th>Safe</th>
<th>Sense of Threat</th>
<th>Alert</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Threat</td>
<td>Safe</td>
<td>Everything is great! <em>Relax</em></td>
<td></td>
</tr>
<tr>
<td>General Danger</td>
<td></td>
<td>People afraid and limiting behavior <em>Communicate</em></td>
<td></td>
</tr>
<tr>
<td>Invisible Danger</td>
<td></td>
<td>People unaware of threat <em>Mitigate danger</em> <em>Communicate with care</em></td>
<td></td>
</tr>
<tr>
<td>Clear Danger</td>
<td></td>
<td>People aware of real danger <em>Mitigate and communicate</em></td>
<td></td>
</tr>
</tbody>
</table>

**Traditional risk analysis**

If this goes wrong, what happens? How do we prevent it? What do we communicate?
### Sense vs. Presence

How do you prioritize user worries versus real dangers?

<table>
<thead>
<tr>
<th>Sense of Threat</th>
<th>Presence of Threat</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Safe</strong></td>
<td>Safe</td>
<td>Mitigate and communicate</td>
</tr>
<tr>
<td></td>
<td>No Threat</td>
<td>Everything is great! Relax</td>
</tr>
<tr>
<td></td>
<td>General Danger</td>
<td>People afraid and limiting behavior Communicate</td>
</tr>
<tr>
<td></td>
<td>Invisible Danger</td>
<td>People unaware of threat Mitigate danger Communicate with care</td>
</tr>
<tr>
<td><strong>Bad</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Clear Danger</td>
<td>People aware of real danger Mitigate and communicate</td>
</tr>
</tbody>
</table>

#### User perception affects action.

When people don’t know, they assume the worst.

If you’re legit, how do people know? How can they verify it?

Fear spreads in a vacuum.

---

**Note:**

- **Sense of Threat** refers to how people perceive the risk.
- **Presence of Threat** refers to the actual risk level.
- **Action** indicates the appropriate response based on the perceived and actual threat levels.
### Sense vs. Presence

How do you prioritize user worries versus real dangers?

<table>
<thead>
<tr>
<th>Presence of Threat</th>
<th>Safe</th>
<th>Alert</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Threat</td>
<td>Safe</td>
<td>Sense of Threat</td>
</tr>
<tr>
<td></td>
<td>Relax</td>
<td>General Danger</td>
</tr>
<tr>
<td></td>
<td>People afraid and limiting behavior, <em>Communicate</em></td>
<td></td>
</tr>
<tr>
<td>Invisible Danger</td>
<td>Mitigate danger, <em>Communicate with care</em></td>
<td></td>
</tr>
<tr>
<td>Clear Danger</td>
<td>People aware of real danger, <em>Mitigate and communicate</em></td>
<td></td>
</tr>
</tbody>
</table>

**Beware of scared PR / legal / etc teams**

One school believes “the less you say, the less risk.”

If people are already afraid, you’re already paying the price.
How do you prioritize user worries versus real dangers?

<table>
<thead>
<tr>
<th>Safe</th>
<th>Sense of Threat</th>
<th>Alert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Safe</td>
<td>No Threat</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Everything is great!</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Relax</em></td>
<td></td>
</tr>
<tr>
<td>Safe</td>
<td>General Danger</td>
<td></td>
</tr>
<tr>
<td></td>
<td>People afraid and limiting behavior</td>
<td>Communicate</td>
</tr>
<tr>
<td>Bad</td>
<td>Invisible Danger</td>
<td></td>
</tr>
<tr>
<td></td>
<td>People unaware of threat</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Mitigate danger</em></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Communicate with care</em></td>
<td></td>
</tr>
<tr>
<td>Bad</td>
<td>Clear Danger</td>
<td></td>
</tr>
<tr>
<td></td>
<td>People aware of real danger</td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>Mitigate and communicate</em></td>
<td></td>
</tr>
</tbody>
</table>

But look for unexpected allies!

PR, legal, marketing, or public policy teams often *want* to communicate.

They often fear the others will say no! Talk to everyone.
Use this to **talk about perception in the language of risk.**

<table>
<thead>
<tr>
<th>Sense of Threat</th>
<th>Presence of Threat</th>
<th>Safe</th>
<th>Alert</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No Threat</strong></td>
<td>Safe</td>
<td>Safe</td>
<td>Alert</td>
</tr>
<tr>
<td>Everything is great! <em>Relax</em></td>
<td>People afraid and limiting behavior <em>Communicate</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Invisible Danger</strong></td>
<td>Bad</td>
<td>Bad</td>
<td>Alert</td>
</tr>
<tr>
<td>People unaware of threat <em>Mitigate danger</em> <em>Communicate with care</em></td>
<td>People aware of real danger <em>Mitigate and communicate</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*If you’re in the red, you have a problem. The risk to your company is real in all three squares!*
You’re worried about a risk; the product VP isn’t. Are they right? Are they uninformed? Do they not care?

Activists view systems through the lens of a single problem. System operators view problems through the lens of a single system.
Activists vs. Operators

The system that most affects the problem ≠ The problem that most affects the system
The system that most affects the problem ≠ The problem that most affects the system
Activists vs. Operators

The system that most affects the problem ≠ The problem that most affects the system
Sometimes priorities don’t align. When this happens, you often want something from an operator.

- You need to build social capital:
  - Do they have a persistent annoyance you can fix?
  - Can your proposal move the metrics they worry about?
  - If you can’t do it, can you get someone else to?

Look for their unspecified fears and pressures – and help fix them!
Caveat: The last bit applies to *non-adversarial* relationships when *you are not the person whose resources are being drained*.

- If basic goals don’t align, this doesn’t work well.
- If you’re the injured party, you don’t have the resources to do this.

But privacy professionals are often in this situation – Use it!
The Moral of the Story

Triage is both about estimating risks and convincing people. The most useful tools give you a common language to do this.

Flox: “Remember the human”:

- People want to feel safe and be safe.
- That includes users and stakeholders alike.
- When you can help them do that, good things can flow.

A. V. Flox, *Disrupting the Bystander*, Thorntree Press, 2019