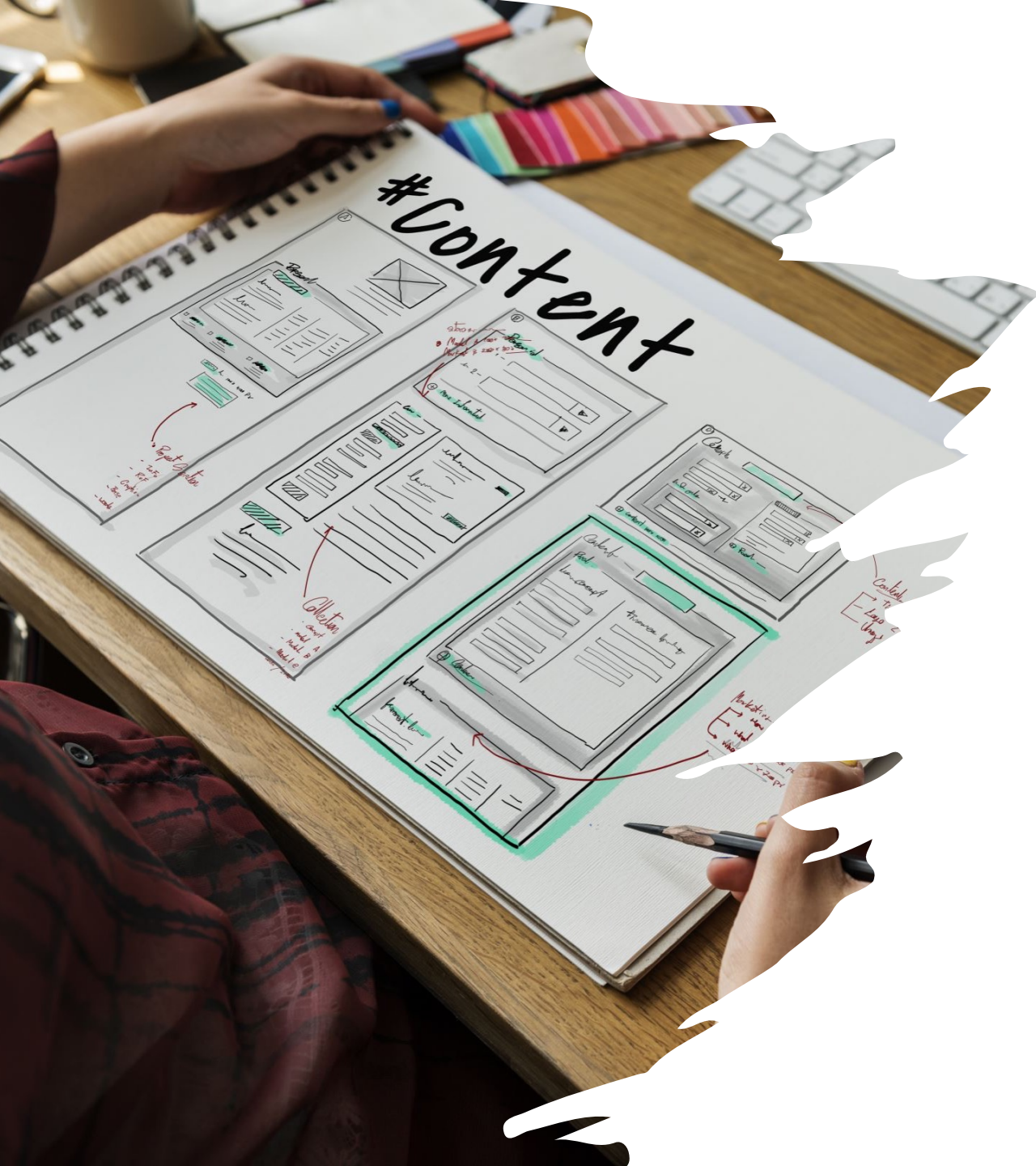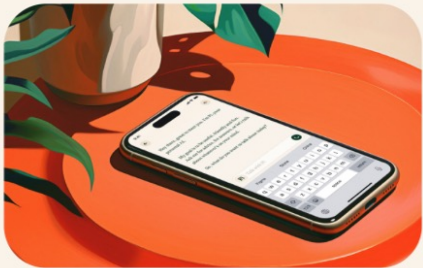# Encoding Human Values

Values in Design

- Design decisions encode values:

  - Efficiency
  - Privacy
  - Beauty
  - Truth
  - Fairness
  - Sustainability

➢ Design decisions express what we care about!

Values in Design: Chatbots

Our first AI is called Pi, for personal intelligence, a supportive and empathetic conversational AI.

I'm

Pi

Your *personal* AI.

- Empathy
- Respect
- Solidarity
- Kindness
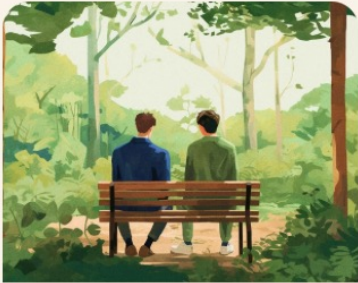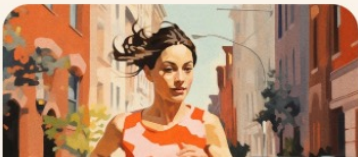- Support

Welcome to Pi

Brainstorm ideas

I need a safe space for something

Relationship advice

Practice a big conversation

Take your time and know that I'm here to listen. What's been happening?

Values in Design: Chatbots

- Empathy
- Respect
- Solidarity
- Kindness
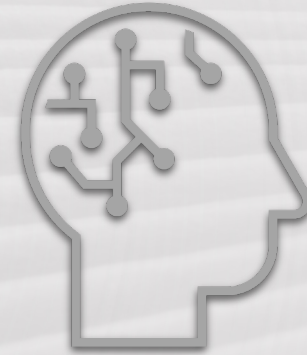- Support

Where do values emerge from?

- Values emerge from:

    - Designer understanding
        - Definition
        - Specification of instrumental design features
    - User perception
    - Broader context

Two Steps

Locating values   Defining values

Locating Values

- Key actors

- Functional Description

- Constraints

- Societal input

Locating Values

Collateral values: values that crop up as side effects of design decisions.

➤ Intentions vs. Impact
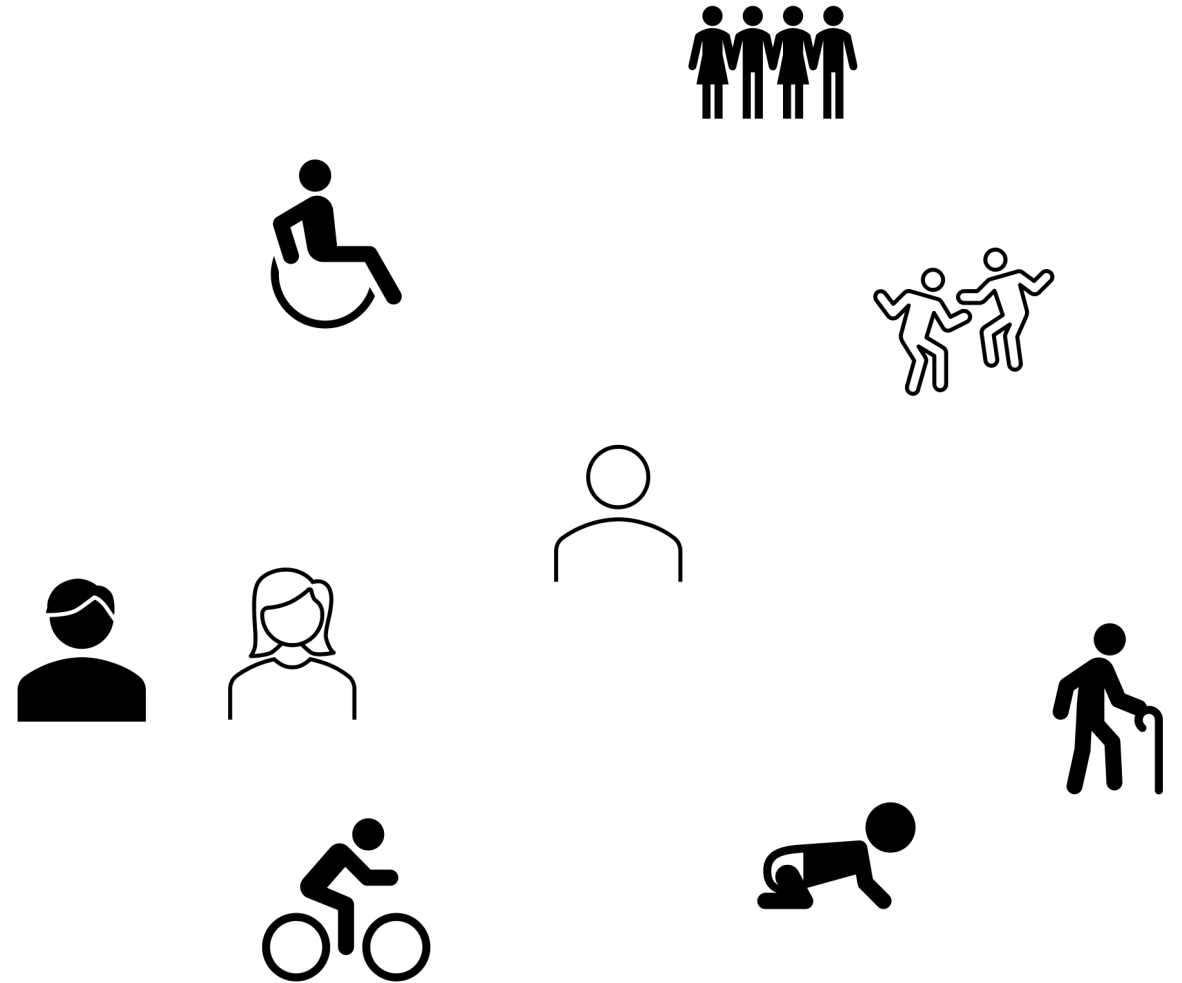
Locating Values

- Default assumptions: Standardization, power, and discrimination

    - *Preexisting bias*, bias that exists in broader society, culture, and/or institutions is reproduced, either intentionally or unintentionally.

    - *Technical bias:* underlying aspect of the technology reproduces bias.

    - *Emergent bias:* a system that may not have been biased given its original context of use or original user base comes to exhibit bias when the context shifts or when new users arrive.
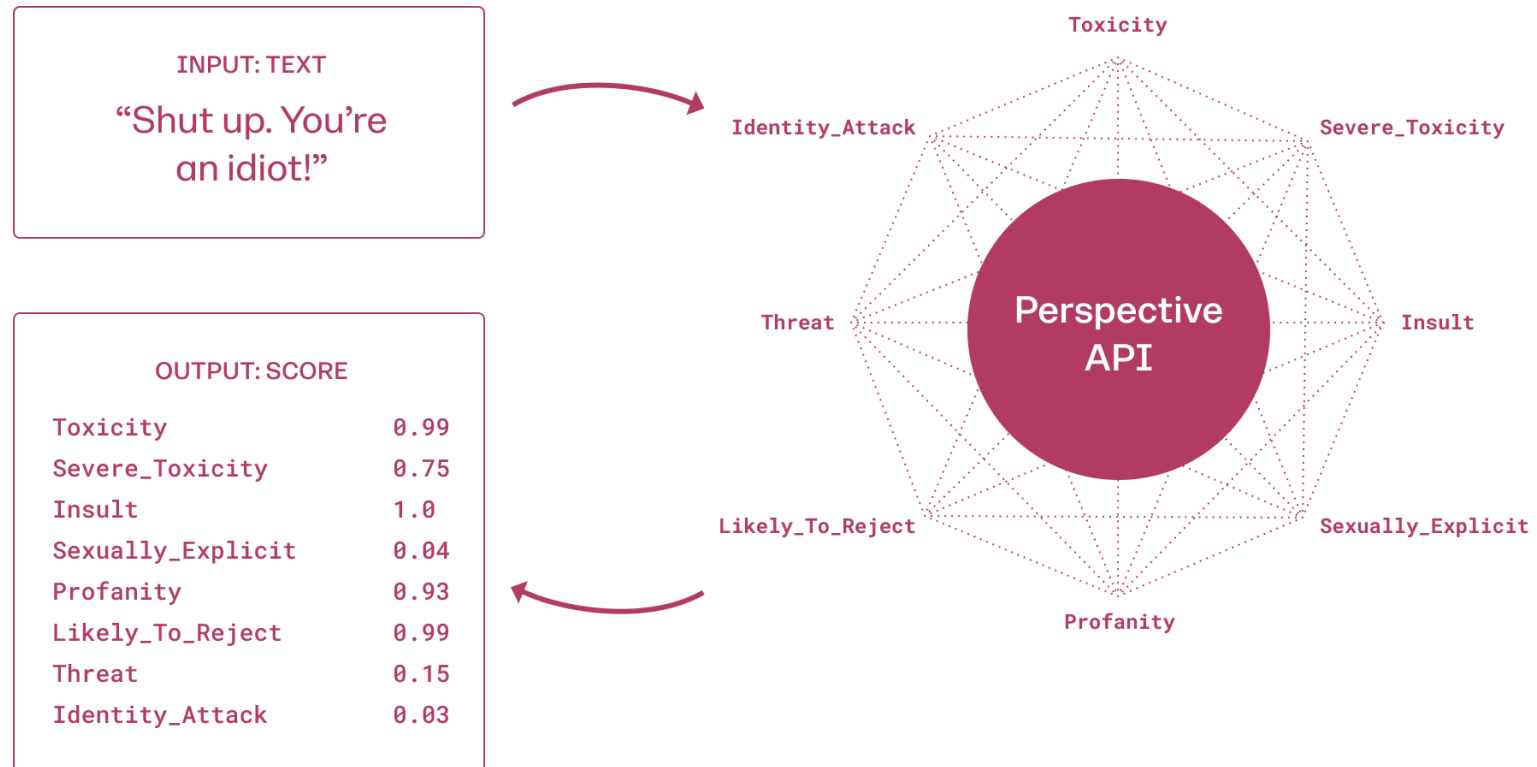
- Standard user
  - Gender
  - Age
  - Ability
  - Race
  - Zip code
  - Access to technology
  - Needs

➢Burden distribution

What counts as "toxicity"?

INPUT: TEXT

"Shut up. You're an idiot!"

OUTPUT: SCORE

Toxicity	0.99
Severe_Toxicity	0.75
Insult	1.0
Sexually_Explicit	0.04
Profanity	0.93
Likely_To_Reject	0.99
Threat	0.15
Identity_Attack	0.03

Toxicity

Identity_Attack

Severe_Toxicity

Threat

Perspective API

Insult

Likely_To_Reject

Sexually_Explicit

Profanity

"Rude, disrespectful, or unreasonable language that is likely to make someone leave a discussion."

Defining Values

- Ethical and political values are abstract, controversial, and difficult to define.

- Definition and analysis connect abstract values to concrete design features.

- Products can miss their mark entirely if values are not well-defined

Defining "inclusion"?

- Welcoming any kind of content and any topic that is of users' interest.

- Protecting vulnerable users from insult and psychological harm by banning certain words and topics.

oft's teenage chatbot, Zo.

Value Conflicts

- Values may conflict with one another.
  - Inevitable consequence of value pluralism (i.e. the recognition that different things matter to us, often to the same degree)

- These conflicts may appear intractable
  - This does not mean that we should throw up our hands
  - Rather, we should strive to make conscientious and responsible choices

Value Conflicts

- Three approaches

    - **Dissolving**: developing a creative alternative that avoids conflict and achieves all values in question.

    - **Compromising**: promoting each of the values in question to a different degree.

    - **Trading-off:** Giving up one or some values in favor of others.