How can we ensure safety in autonomous systems that operate with people in the real-world?
How can we ensure safety in autonomous systems that operate with people in the real-world?
Emergence of Autonomy in Planes

Fatalities per million departures

Time


Today’s Roadmap

- Introduction
- Levels of Autonomy and ADAS
- Value Sensitive Design and Empirical Evaluation
- Human Metrics
## SAE Levels for Autonomy

<table>
<thead>
<tr>
<th>Level 0</th>
<th>Provide warnings with no interventions</th>
<th>Driving</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>Partial or shared control between the driver and the vehicle</td>
<td>Hands on</td>
</tr>
<tr>
<td>Level 2</td>
<td>Performs all driving functions, but may require quick intervention from the driver</td>
<td>Hands off</td>
</tr>
<tr>
<td>Level 3</td>
<td>Performs all driving functions, but may call on the driver for assistance with timed handoffs</td>
<td>Eyes off</td>
</tr>
<tr>
<td>Level 4</td>
<td>Fully autonomous, but restricted to certain areas or conditions with prescribed handoffs</td>
<td>Mind off</td>
</tr>
<tr>
<td>Level 5</td>
<td>Fully autonomous with no restrictions</td>
<td>Steering wheel optional</td>
</tr>
</tbody>
</table>

[https://www.sae.org/standards/content/j3016_201806/](https://www.sae.org/standards/content/j3016_201806/)
Advanced Driver Assistance Systems

- **Forward Collision Mitigation**
  - Warning: ↓27%
  - Autobrake: ↓50%
  - Rear-end rate: ↑20%*

- **Lane Keeping Assistance**
  - Warning: ↓11%

- **Blind Spot Detection**
  - Warning: ↓14%

- **Rearview Assistance**
  - Camera: ↓17%
  - Warning: ↓22%
  - Autobrake: ↓78%

Image Credit: Google Image Search

https://www.iihs.org/topics/advanced-driver-assistance
Today’s Roadmap

Introduction

Levels of Autonomy and ADAS

Value Sensitive Design and Empirical Evaluation

Human Metrics
Value Sensitive Design

• How do designers connect human values to engineering specification?
  • Explicit encoding of safety and legal constraints
  • Ex: safety and efficiency / comfort → how to handle conflicting values?
Value Sensitive Design: Ethical Considerations

- Philosophy formally characterizes ethical entanglements
- **Deontology**: follow set of rules with no exception
- **Consequentialism (utilitarianism)**: cost-based framework that evaluates choices based on consequences (considers outcome of chosen action)

How to incorporate into engineering design?
- Actions bounded by deontological constraints
- Cost functions realized via consequentialism

\[
\begin{align*}
\text{minimize} & \quad \text{path deviation} \\
\text{action} & \quad \text{steering rate} \\
& \quad \text{traffic law violation}^* \\
\text{subject to} & \quad \text{actuator limits} \\
& \quad \text{obstacle avoidance} \\
& \quad \text{traffic law violation}^* \\
\end{align*}
\]

\{ consequential costs \}
\{ deontological constraints \}

Value Sensitive Design

• How do designers connect human values to engineering specification?
  • Explicit encoding of safety and legal constraints
  • Ex: safety and efficiency / comfort → how to handle conflicting values?

• VSD iterates over conceptual, technical, and empirical investigations by connecting with stakeholders
  • Conceptualization: Who are the direct and indirect stakeholders? What are the human values?
  • Technical Implementation: What are the assumptions? What are the inputs and goals? What is the formal approach? How does it encode human values?
  • Empirical Analysis: How will you design the experiment? What is the formal hypothesis? What is the baseline?
  • Iterate: Creates documentation and justification for design decisions
# Case Study: Pedestrian Collision Avoidance

## Direct and Indirect Stakeholders
- Vehicle Occupants
- Traffic Authorities
- Pedestrians
- Surrounding vehicles
- Bystanders
- Community and society

## Human Values
- Safety, Care / Respect for others → safety
- Legality, Respect for authority → legality
- Mobility, Autonomy → efficiency
- Mobility, Trust and transparency → smoothness
- Fairness and reciprocity → stakeholder

## Technical Considerations
- Safety constraints: Requires distance, velocity, detection
- Legality: Vehicle codes
- Time efficiency: Requires vehicle speed
- Smoothness: Requires acceleration

## Problem Statement
**Given:** Distance to crosswalk  
Vehicle speed  
Detection of pedestrian crossing  

**Control input:** Longitudinal acceleration  

**Goal:** Drive safely, efficiently, and smoothly

---

Empirical Evaluation: Experimental Design

**Conditions (Treatments)**
- Drug vs placebo
- Random vs optimal
- RRT* vs RL

**Measures (Responses)**
- Symptom progression
- Comfort
- Cost

**Experimental Components**

**Experimental Units (Subjects)**
- Patients
- Users
- Domains / Problems

**Assignment Method (Allocation)**
- Random
- Comparative (subjects “see” both)
Conditions, Measures, and Hypotheses

• **Independent Variables:** conditions
  - What are you manipulating? (e.g., drug type, controller type, detection method)
  - If there are $n$ possible values (levels) for one IV, then we run $n$ experiment settings
  - If there are two IVs with $n$ and $m$ levels, we run $nm$ experiment settings

• **Dependent Variables:** measured outcomes
  - What are you measuring? (e.g., symptoms, trust, accuracy)
  - Check for construct validity (e.g., does IQ measure intelligence? Do user ratings of predictability indicate predictability?)

• **Hypothesis:** Independent variable $x$ (positively) affects dependent variable $y$

• **Type of Experiment**
  - Observational, prevalence (cross-sectional), and controlled
  - *Exploratory Data Analysis:* What the data can tell us beyond the formal modeling or hypothesis testing?
Case Study: Kidney Stone Treatments

• A study was performed to compare the effectiveness of two classes of treatments for kidney stones

<table>
<thead>
<tr>
<th>Stone Size</th>
<th>Open Surgery</th>
<th>Percutaneous Nephrolithotomy (PN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 2cm</td>
<td>93% (81/87)</td>
<td>87% (234/270)</td>
</tr>
<tr>
<td>&gt;= 2cm</td>
<td>73% (192/263)</td>
<td>69% (55/80)</td>
</tr>
<tr>
<td>Overall</td>
<td>78% (273/350)</td>
<td>83% (289/350)</td>
</tr>
</tbody>
</table>

• PN has a higher success rate overall, but open surgery has a higher success rate in both group

• The assumption is that the decision for the treatment is independent on the size of the stone – which is incorrect!

→ Simpson’s paradox // “correlation does not equal causation”

http://courses.ieor.berkeley.edu/ieor165/lecture_notes/ieor165_lec15.pdf
Causality and Confounds

Confounds: Variable whose effect cannot be distinguished from the effect of the tested conditions (independent variable).

Tools to eliminate confounds?
- Randomization
- Within subjects / counter-balanced
- Block strategies
- Be the devil’s advocate
  - Optimize the baseline
  - Use measures that do due justice to the baseline
Value Sensitive Design

• How do designers connect human values to engineering specification?
  • Explicit encoding of safety and legal constraints
  • Ex: safety and efficiency / comfort → how to handle conflicting values?

• VSD iterates over conceptual, technical, and empirical investigations by connecting with stakeholders
  • Conceptualization: Who are the direct and indirect stakeholders? What are the human values?
  • Technical Implementation: What are the assumptions? What are the inputs and goals? What is the formal approach? How does it encode human values?
  • Empirical Analysis: How will you design the experiment? What is the formal hypothesis? What is the baseline?
  • Iterate: Creates documentation and justification for design decisions
Today’s Roadmap

Introduction

Levels of Autonomy and ADAS

Value Sensitive Design and Empirical Evaluation

Human Metrics
Common Metrics for HRI: Biasing Effects

• Communication: delay, jitter, bandwidth
• Robot Response: timing factors, reliability
• User / Human Factors:
  – Performance shaping factors
    • Operational: tactics, time constraints
    • Equipment: workspace layout
    • Task: complexity, repetitiveness
    • Personnel: training, motivation, stress
    • External: visibility
  – Interface (design factors, clutter, input mechanism)
  – Role (supervisor, operator, mechanic, peer, bystander)

Common Metrics for HRI: (Shared) Metrics

• System Performance
  ▪ Quantitative (effectiveness, efficiency)
  ▪ Subjective Ratings (incorporate stakeholders for feedback)
  ▪ Utilization of Human-Robot teaming (requests for assistance)

• Operator Performance
  ▪ Situational Awareness → [tool] Situational Awareness Glof Technique (SAGAT)
  ▪ Workload → [tool] NASA-Task Load IndeX (NASA-TLX)
  ▪ Accuracy of mental models (conceptual, movement, spatial, and modality compatibility)

• Robot Performance
  ▪ Self-awareness (assessment, monitoring, understanding of intrinsic limitations)
  ▪ Human Awareness (ability to model and predict, user-oriented perception, adaptation)
  ▪ Autonomy (functionality when encountering anomalies, no human help (neglect tolerance))

Sample Trust Subjective Measures

The following statements were used to assess general propensity to trust:
1. I am suspicious of the [system name]’s intent, action, or outputs.
2. The [system name]’s actions will have a harmful or injurious outcome.
3. The [system name] is dependable.
4. The [system name] is reliable.
5. I can trust the [system name].

Questions from Muir’s Questionnaire include:
1. To what extent can the [system name] behavior be predicted?
2. To what extent can you count on the [system name] to do its job?
3. What degree of faith do you have that the robot will be able to cope with similar situations in the future?
4. Overall, how much do you trust the [system name]?

Case Study: Trust in ADAS

- Participants drove a 2016 Toyota Prius, 2016 Honda Civic, 2017 Audi Q7, or 2016 Infiniti QX60 for several weeks
  - Equipped with forward collision warning, adaptive cruise control, active lane keeping, side-view assist, and lane departure warning
- Subjects reported mileage and warnings daily, and trust survey post-use
  - Responses were averaged to create a composite measure of trust ranging from -2 (strongly disagree) to +2 (strongly agree)

COMPOSITE TRUST IN ADAS

What does “trust” mean?

Trust is a willingness to take risk and to be vulnerable to another party.

One proxy for assessing trust is to gauge whether users rely on their active safety systems.

Is ADAS used?

Observations at dealerships of seven automakers found:
• 93% of forward collision systems were activated
• 99% of the blind spot detection and alert systems were activated
• 90% of driver monitoring alerts were activated
• 52% of lane departure warning and lane-keep assists were activated

Which systems do you think people trusted?
Is ADAS used?

Recall that lane departure systems are much less activated by drivers. Users note that lane departure alerts are “more annoying” than forward collision alerts. Why?

<table>
<thead>
<tr>
<th>Is it the alert itself?</th>
<th>Is it a functional aspect of the system?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode of communication and interaction is a common study in Human Factors and Human-Computer Interaction</td>
<td>1. Drivers who do not use a signal deem warnings as false alarms</td>
</tr>
<tr>
<td>→ generally vibration warnings are often “less annoying” than audio</td>
<td>2. Intentional deviations may be required, leads to needless alerts</td>
</tr>
<tr>
<td></td>
<td>3. Cameras may miss-detect lanes, leading to false alarms or missed warnings</td>
</tr>
<tr>
<td></td>
<td>→ False or unnecessary alerts decrease user acceptance</td>
</tr>
</tbody>
</table>

Perceived **usefulness** and **ease of use** are factors that largely determine user acceptance and trust.

Data shows that 40% and 27% of Volvo and Toyota owners, respectively, have experienced unnecessary alerts. 17% and 25% of these owners reported missed warnings.

Many Facets of and Considerations for Trust

**Context and Impact**

- **Relationship between Decision Impact, Trustworthiness and Trust**
  - Trust $T = f(\text{Competence}, \text{Integrity})$
  - Impact of decision:
    - Life-threatening
    - Life-changing
    - Inconvenient
    - Frustrating
  - Trustworthiness: Threshold for trust-based decisions

**Under- and Over- Trust**

- **Expected Loss**: $EL = \sum_{i=1}^{n} \frac{\text{Expected Loss at trial}}{\text{Number of Overrides}}$

**Physiological Response**

- Steering Wheel
- GSR
- PPG Clip

---

Today’s Roadmap

Introduction

Levels of Autonomy and ADAS

Value Sensitive Design and Empirical Evaluation

Human Metrics
References and Other Resources


• **Value Sensitive Design:**

• **Introduction to Human Factors:** [textbook] Handbook of human factors and ergonomics by G. Salvendy

• **Causality:** [textbook] Causality: Models, Reasoning, and Inference by J. Pearl

• **Models and Techniques for Assessing Trust:**
Fatal rate by year - valid end 2009

Fatal per million departures

1st generation: Early jet
2nd generation: 2nd jet generation
3rd generation: Glass-cockpit, Nav display, FMS
4th generation: FBW, Flight Envelope Protection

Sources: Ascend, Airbus
Case Study: Antilock Braking

• Antilock brakes allow for shorter stopping distances and the ability to steer / maintain control while hard braking, especially on wet and slippery surfaces

• Owners of late model cars equipped with antilock brakes were surveyed in North Carolina and Wisconsin regarding their experiences with antilocks

• The survey results indicated that more than 50% of the drivers in North Carolina and 40% in Wisconsin incorrectly indicated how to brake a car in an emergency situation on wet and slippery pavements in a way that will effectively activate the antilock feature

• More drivers in Wisconsin than in North Carolina reported that their cars' antilock feature had been used, but more than 33% of the Wisconsin drivers and 62% of North Carolina drivers said they had never used the antilock feature of their cars' brakes

• Conclusion: people did not understand the automation!

https://www.iihs.org/topics/advanced-driver-assistance