CS-639 — Interaction Design Studio

Ethics & Responsible AI Design*

Professor Bilge Mutlu

* Visuals obtained from the web, used for educational purposes without modification. Attribution on last slide.

Today

— **A2 submitted this morning** — congratulations on completing your intelligent system design

— This week shifts from "how to design" to **"what should we design"**

— W08-W11 gave you the tools. W12 systematized them. W13 asks: **whose values did your design serve?**

— A3 (Ethical Audit & Redesign) introduced today

P1-P12 are prescriptive — do this. Value Sensitive Design is generative — ask this, then decide.

From Design to Responsibility

Phase	Weeks	Question
Foundations	W01-W07	Can you design well?
Intelligence	W08-W11	Can you design with intelligence?
Systems	W12	Can you systematize your design?
Responsibility	W13-W14	Can you design responsibly?

Each phase builds on the last. You cannot design responsibly if you cannot design well — but designing well is not enough.

Part 1: Why Ethics for Intelligent Systems?

Intelligence amplifies ethical stakes

Intelligence Amplifies Stakes

Every material property creates new ethical risks:

Material Property	What It Enables	What Could Go Wrong
Agency (W08)	System acts on user's behalf	Wrong action causes real harm
Proactivity (W09)	System initiates at the right moment	Unwanted interruption, manipulation
Collaboration (W10)	Shared control between user and system	Unclear accountability for outcomes
Context-awareness (W11)	System senses and adapts to situation	Surveillance, privacy violation

The same properties that make intelligent systems powerful make them dangerous.

Agency Harm: Boeing 737 MAX

MCAS overrode pilot input — 346 deaths (2018-2019)

— The system had too much agency and insufficient override controls

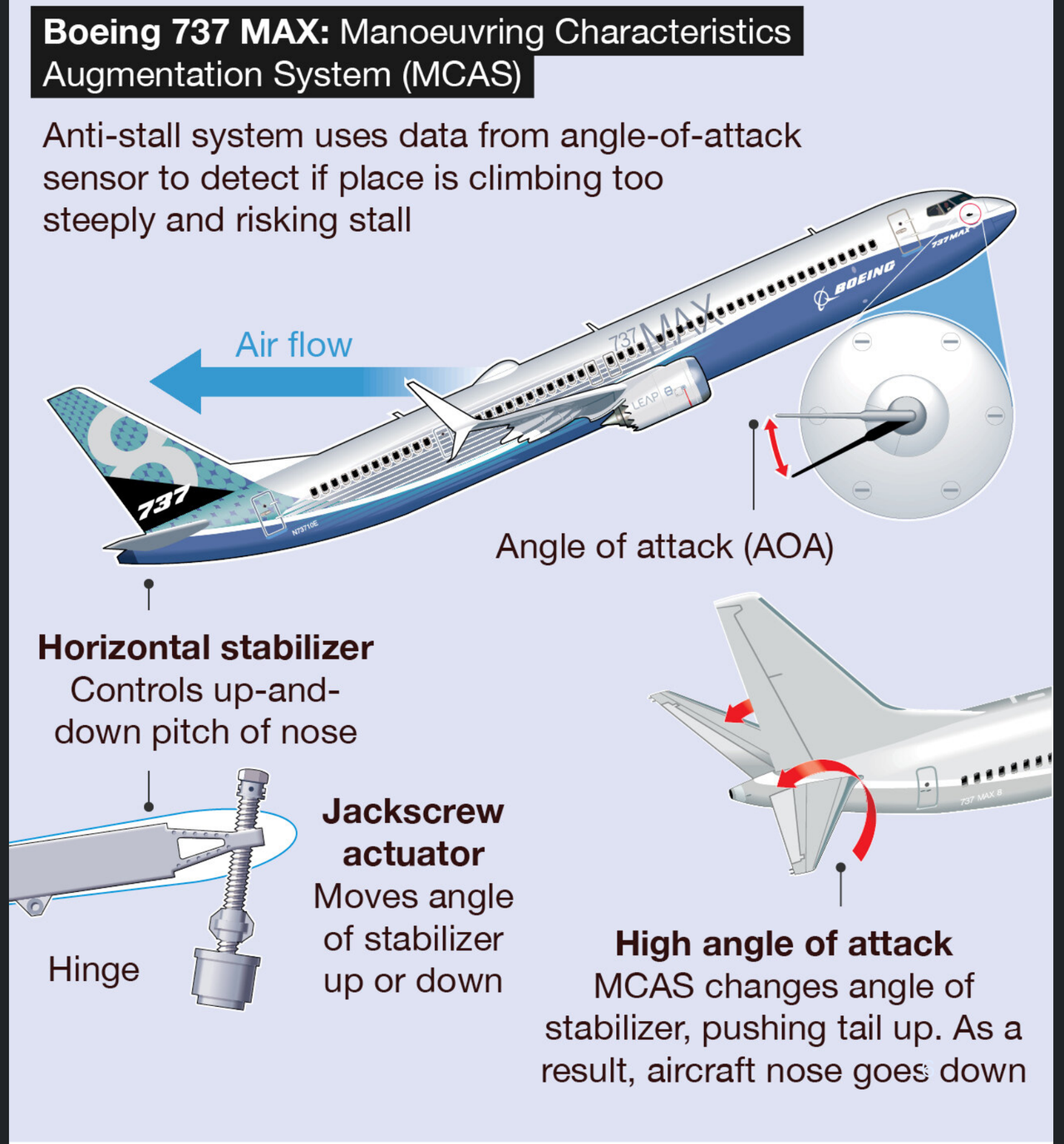
— Pilots couldn't see what the system was doing **(low O)**

— Pilots couldn't predict its behavior **(low P)**

— Pilots couldn't override it quickly enough **(low D)**

OPD failure across the board — a design failure, not just an engineering one.

When agency is high and OPD is low, the system becomes the most dangerous stakeholder.



Proactivity Harm: Zillow iBuyer

Algorithm proactively bought 7,000 homes — $900M+ loss (2021)

— **Agency (W08):** Level 9+ automation, no human approval per purchase

— **Proactivity (W09):** Acted when it should have Asked — stakes too high for auto-action

— **Context (W11):** SA-3 projection failed — predicted prices would keep rising

— **Collaboration (W10):** AI-led, no human-in-the-loop

Every material property failed simultaneously. High agency + proactive action + wrong projection + no human check = catastrophic harm.

Collaboration Harm: Uber Self-Driving Car

Neither side knew who was responsible. A pedestrian died (2018)

— A collaborative handoff failure — neither side knew who was responsible

— The safety driver trusted the system **(assumed AI-led)**

— The system expected human monitoring **(assumed human-led)**

— No indicator showed who was "driving" at any moment

When leadership is ambiguous, no one is in charge. That is when people get hurt.

Context-Awareness Harm:
Strava Heatmap

Aggregated fitness data revealed military bases (2018)

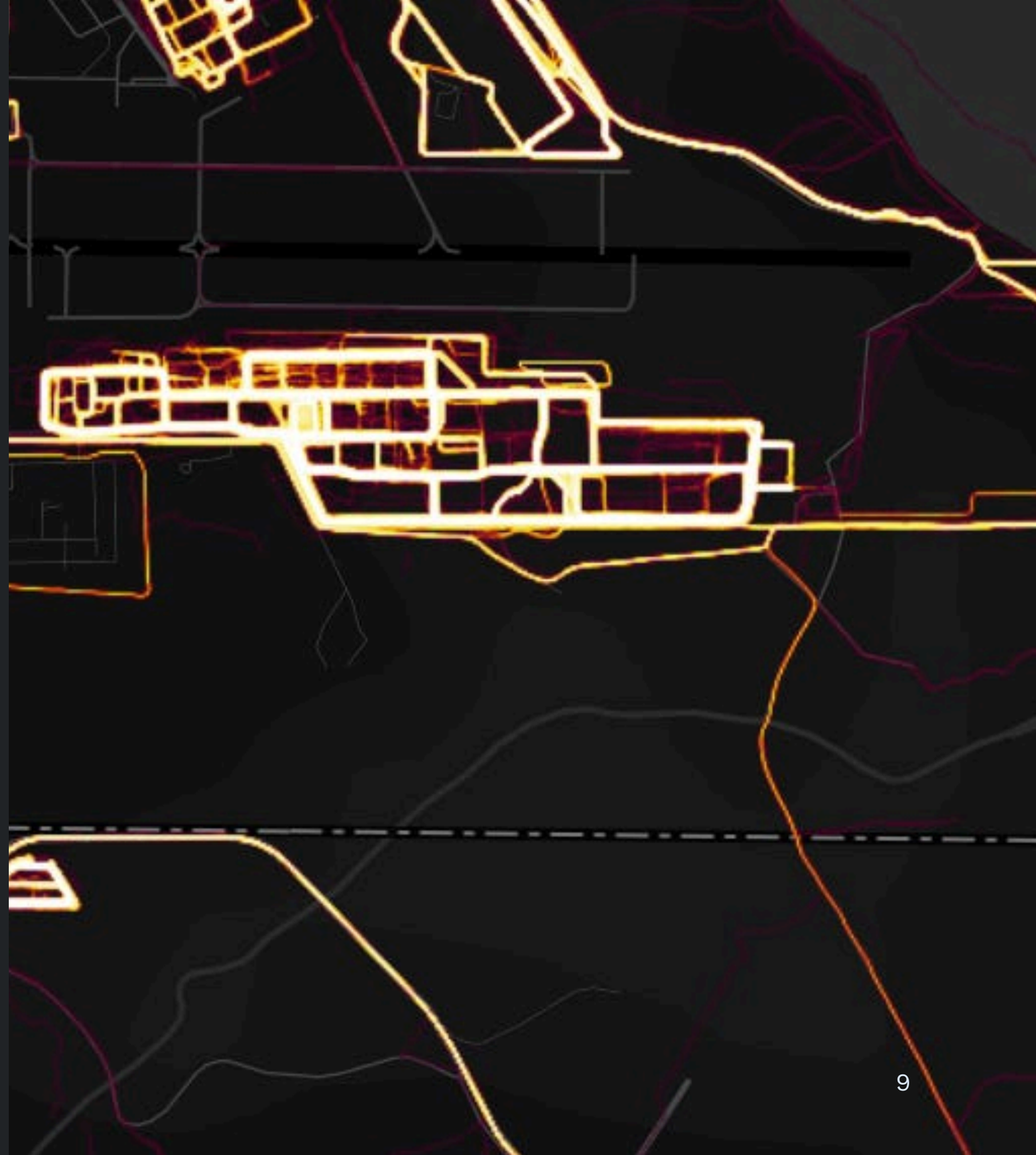
— **SA-1:** GPS, pace, time — harmless raw data

— **SA-2:** "This is a running route" — still harmless

— **SA-3:** Aggregated heatmap revealed military bases with predictable patrol routes

— **Contextual integrity:** route shared with fitness app (appropriate) → public heatmap (violation)

The harm escalated with the SA level. Individual data was benign. Aggregated data created new knowledge no user consented to reveal.



Your A2 System — What Could Go Wrong?

You designed a high-stakes system with all four material properties. Now ask:

— **Who benefits** from your design?

— **Who could be harmed** — even if the system works exactly as intended?

Quick think (1 min): Name one way your A2 system could cause harm if it worked perfectly — not a bug, but a feature that hurts someone.

The most dangerous failures are not bugs. They are features that work perfectly for the wrong people.

Part 2: Value Sensitive Design

A method for surfacing and negotiating values in design

Value Sensitive Design (VSD)

Friedman & Hendry (2019) — Value Sensitive Design: Shaping Technology with Moral Imagination

— VSD is a **method**, not a checklist of correct values

— Core question: whose values are at stake, and how do we surface and negotiate them?

— Three investigations: **conceptual**, **empirical**, **technical**

— Used across HCI, AI ethics, and policy for over two decades

VSD does not tell you what is right. It gives you a method to ask the question — and take the answer seriously.

VSD in Practice: Envisioning Cards

One of VSD's most widely used tools: **Envisioning Cards** (Friedman & Hendry, 2012)

— A deck of prompt cards organized by four themes: **Stakeholders**, **Time**, **Values**, **Pervasiveness**

— Each card poses a design question and suggests an activity

— Used as a reflection tool during design — not prescriptive, but generative

— Our Intelligence Design Pattern Cards from W12 were inspired by this format

VSD works through structured reflection, not checklists. The cards are one way to make that reflection concrete and actionable.

Three VSD Investigations

Investigation	Original Method	A3 Adaptation
Conceptual	Philosophical analysis of stakeholders and values	Stakeholder map
Empirical	User studies, interviews, observation	Failure scenario
Technical	Analysis of what the design enables or constrains	Redesign

The Intelligence Design Pattern Cards from W12 were inspired by VSD's Envisioning Cards. Now you are using the full method.

Conceptual Investigation — Stakeholders

- **Direct stakeholders:** users who interact with the system
- **Indirect stakeholders:** people affected by the system but who do not use it (caregivers, employers, bystanders)
- **Excluded stakeholders:** people the system was not designed for but who encounter it
- For each group: what values matter to them? (privacy, autonomy, fairness, safety, trust)

The stakeholders you did not think about are usually the ones most at risk.

Conceptual Investigation — Values in Tension

Values often conflict — this is normal, not a design failure:

— **Safety vs. autonomy** — the system restricts freedom to prevent harm
— **Personalization vs. privacy** — better service requires more data
— **Efficiency vs. transparency** — explaining slows things down
— **Fairness vs. accuracy** — optimizing for one group may disadvantage another

Good design does not eliminate tensions. It makes them visible and negotiates them deliberately.

Part 3: Five Ethical Principles

A global consensus on what ethical AI requires

Global AI Ethics Guidelines

Jobin et al. (2019) A survey of **84 AI ethics guidelines** from governments, companies, research institutions, NGOs and found 11 recurring principles, with five appearing in over half of all guidelines:

Principle	Prevalence	Course Connection
Transparency	73/84	P3, P7, explainers (W12)
Justice & Fairness	68/84	Bias, inclusion, "who is the default user?"
Non-maleficence	60/84	Failure scenarios, harm prevention
Responsibility	60/84	Accountability, collaboration handoffs (W10)
Privacy	47/84	Contextual integrity (W11)

These are what the world's AI ethics community converges on and map directly to what you have been learning all semester.

1. Transparency

Can users see what the system does and why? (73/84 guidelines)

— Connects to: P3, P7, explainer components from W12

— Levels: what data is collected → what inference is made → what action is taken → why

— Example: A credit scoring app that denies a loan but will not say why vs. one that shows the contributing factors

Transparency is not just showing data. It is showing reasoning at a level the user can act on.

2. Justice & Fairness

Does the system treat all users equitably? (68/84 guidelines)

— Bias sources: training data, feature selection, optimization targets, edge cases

— Who benefits from the default? Who is the "average user" the system was designed for?

— Example: Voice assistants that work well for native English speakers but fail for accented speech

Fairness is not treating everyone the same. It is ensuring the system works for everyone it affects.

3. Non-maleficence

Does the system avoid causing harm? (60/84 guidelines)

— Harm includes psychological, financial, social, and dignitary — not just physical

— Connects to: override controls (W12), failure paths (A2), Act/Ask/Wait (W09)

— **Negative:** Navigation app routes through danger to save 3 min

— **Positive:** Waze "avoid difficult intersections" — constrains routes for safety

Non-maleficence is not "do no harm" — it is "understand the harm your design makes possible, and design against it."

4. Responsibility

Who is accountable when the system causes harm? (60/84 guidelines)

— Connects to: collaboration (W10) — shared control makes accountability murky

— Boeing and Uber (Part 1) are responsibility failures — no one was clearly accountable

— Responsibility includes: acting with integrity, attributing liability, enabling redress

— **Positive:** GitHub Copilot checks code against public repos and shows attribution — accountability built into the product

Responsibility is not about blame. It is about designing systems where someone can always answer: "Who is accountable for this outcome?"

5. Privacy

Does the system respect information norms? (47/84 guidelines)

— Revisit Barth et al. (2006) from W11 — now the central ethical frame
— Two norms: **appropriateness** (is this data fitting here?) and **flow** (who receives it?)
— Does the system create **new information** through inference the user never consented to share?
— Example: Spotify Wrapped reveals private listening habits publicly — context shifts from personal to social

The system did not just collect data. It created new knowledge the user never consented to share.

Part 4: The Ethical Audit

A structured method for examining your A2 system

Ethical Audit Method

Six steps — this is the method you will use for A3:

1. **Stakeholder map** — direct, indirect, excluded
2. **Values inventory** — what matters to each stakeholder?
3. **Tension identification** — where do values conflict?
4. **Five-principle check** — transparency, fairness, non-maleficence, responsibility, privacy (Jobin et al., 2019)
5. **Failure scenario** — write a concrete story where the tension causes harm
6. **Redesign** — change the design to address the tension

This is the method. A3 asks you to apply it to your A2 system.

Running Example — Diabetes App Ethical Audit

Applying the method to W11's running example:

Step	Application
Stakeholders	Patient (direct), caregiver (indirect), insurer (indirect), employer (excluded)
Tension	System collects glucose data to help patient, but insurer could use it to raise premiums
Dimensions	Privacy (contextual integrity violation), Autonomy (patient cannot control data flow)
Failure	Insurer accesses glucose patterns, denies coverage for "pre-existing condition"
Redesign	Data stays on-device; patient explicitly controls sharing per-recipient

Your A3 audit follows this same structure — applied to your A2 system.

Quick Check — Your A2 System

In pairs (2 min):

1. Name one **indirect stakeholder** of your A2 system
2. Name one **value tension** (e.g., safety vs. autonomy)
3. Name one way **contextual integrity** could be violated

Share out 2-3 pairs.

This is the starting point for A3. The tensions you identify now become the core of your audit.

This Week

A3 introduction and ethical audit

Assignment 3: Ethical Audit & Redesign

— **Due Monday, May 5 | 15% of grade**

— Audit your A2 system using VSD (Friedman & Hendry, 2019)

— Identify one ethical tension, write a concrete failure scenario, redesign to address it

— **Deliverables:** stakeholder map, values inventory, failure scenario, redesigned screens, critical reflection (1000-1500 words)

— Builds directly on A2 — no new system, just critical examination

You built an intelligent system. Now stress-test it.

Reflection: Ethical Audit Start

Due before next Monday (Apr 27) | Graded: check system

Three parts:

1. **Stakeholder map** — identify direct, indirect, and excluded stakeholders of your A2 system
2. **One ethical tension** — name a specific conflict between stakeholder values
3. **Draft failure scenario** — write 2-3 sentences describing a concrete situation where the tension causes harm

Submit on Canvas.

This is the first step of A3. Start here, and the rest of the assignment follows.

Before Next Week

— **Submit reflection** before Monday, April 27

— **Begin A3 ethical audit** — use the six-step method from today

— **Friday guest lecture (Apr 24):** Professor Jodi Forlizzi (CMU) — "The Role of Design in the Age of AI." Attend during class time. **W14 reflection will be based on this talk.**

— **Optional reading:** Friedman & Hendry (2019), Chapters 1-2 — VSD foundations and method

A3 is lighter than A2 — no new system, just a critical lens on what you already built. Start with the stakeholder map and let the tensions emerge.

References

Ethics & VSD:

— Jobin et al. (2019). "The Global Landscape of AI Ethics Guidelines" — Nature Machine Intelligence

— Friedman & Hendry (2019). Value Sensitive Design: Shaping Technology with Moral Imagination — MIT Press

— Friedman & Hendry (2012). "The Envisioning Cards" — CHI '12

— Barth et al. (2006). "Privacy and Contextual Integrity" — IEEE S&P

— Amershi et al. (2019). "Guidelines for Human-AI Interaction" — CHI '19

Course Frameworks:

— Parasuraman et al. (2000). "A Model for Types and Levels of Human Interaction with Automation" — IEEE SMC

— Horvitz (1999). "Principles of Mixed-Initiative User Interfaces" — CHI '99

— Johnson et al. (2014). "Coactive Design" — JCEDM

— Dey (2001). "Understanding and Using Context" — Personal and Ubiquitous Computing

— Jiang et al. (2023). "A Situation Awareness Perspective on Human-AI Interaction" — IJHCI

Media Sources

Boeing 737 MAX / MCAS | Zillow iBuyer | Uber Self-Driving Car | Strava Heatmap | GitHub Copilot | Waze | Spotify Wrapped | Envisioning Cards