Responsible AI for Business

Savneet Singh
5-virtues.com

https://www.linkedin.com/in/savneetsingh1/

What's ahead

- **First session- AI Ethics in Practice & Legal Frameworks**
- Privacy, Surveillance & AI
- AI & Intellectual Property Wars
- Building Responsible AI Practices

Why do we care?

- Why AI ethics matters for employees, professionals and tech companies
- How this applies to sales AI systems specifically

discrimination lawsuits

privacy violations

regulatory penalties

professional sanctions

workplace surveillance

customer privacy

automated decision-making

proprietary client information

The SHIFT Framework

S - Safe AI
Protecting people and systems from unintended harm

H - Humane AI
Innovate and progress with human dignity and wellbeing at every step

I - Innovative

Creating meaningful progress that benefits society

F - Fair AI
Equitable outcomes and eliminating bias

T - Transparent

Operating with complete transparency and accountability

Safe and Secure

- C.I.A. triad: Confidentiality, Integrity, and Availability
- Data Governance

- Existential Risks in AI
- Unintended Consequences
- Value Alignment Challenges
- Lack of Transparency
- Biases

Bias

- What is bias?
- Why is understanding bias important?
- Over 100 types of cognitive biases exist
- Critical to audit machine learning models

Oxford dictionary defines bias as "a strong feeling in favor of or against one group of people, or one side in an argument, often not based on fair judgment".

Humane AI

-Trustworthy AI makes sure that systems are ethical by design.

-Responsible AI makes sure that humans remain in control.

-Human centric AI makes sure that technology serves humanity’s best future

Innovate, AI for good

Transparency and Explainability

The "Black Box" Problem

When we talk about "explainability," we're focusing on answering questions like:

- Why did the AI system make this decision?
- What factors influenced its choice?
- How does the system arrive at its conclusions?

Key Characteristics

- Explainability is focused on making AI systems understandable to humans
- It aims to provide transparency into the "black box" nature of complex AI models
- It helps build trust between users and AI systems
- It addresses both technical and non-technical audiences' needs for understanding

Bias in the Machine Learning Lifecycle

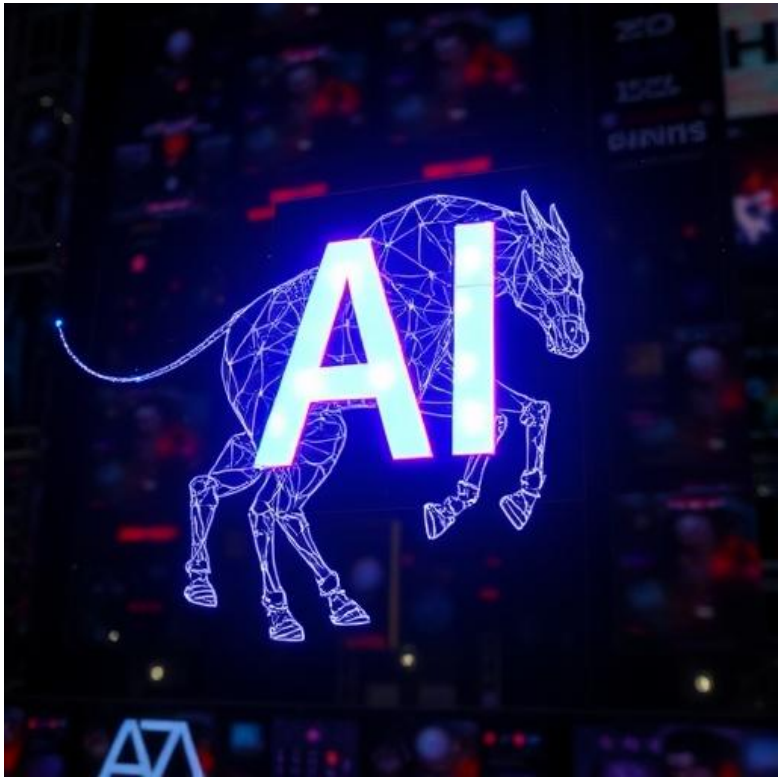
Data Collection

Data Preparation

Model Training

Model Evaluation

Model Deployment



What LLM Users Must Watch For

Data Exposure

Prompt Injection

Data Poisoning

Financial Impact

What can you do right now?

- Verify AI outputs (don't blindly trust generated pitches).
- Demand transparency from your AI
- Push back on unethical features
- Advocate for training on your company's AI ethics policies.

QA

Feel free to reach out!

Data Privacy

Session 2

Data privacy refers to the right of individuals to control how their personal information is collected, used, shared, and protected. It encompasses the principles, policies, and practices that safeguard sensitive data from unauthorized access, misuse, or exploitation.

Key Components of Data Privacy

Personal Data

- Identifiers: Name, ID number, location data.
- Sensitive data: Health records, biometrics, political opinions, sexual orientation, financial details.
- Digital footprints: IP addresses, browsing history, social media activity

Key Components of Data Privacy

Core Principles

- **Consent**: Individuals must knowingly agree to data collection/use.
- **Purpose Limitation**: Data should only be used for specified, legitimate purposes.
- **Minimization**: Collect only data essential for the stated purpose.
- **Accuracy:** Data must be kept up-to-date and correct.
- **Security**: Protect data from breaches or leaks.
- **Transparency**: Clear communication about how data is handled.
- **Accountability**: Organizations must demonstrate compliance and address violations.

Why It Matters

- **Autonomy: P**eople should control their digital identity.
- **Safety:** Prevents harm like discrimination, fraud, or stalking.
- **Trust:** Essential for ethical relationships between users and organizations.
- **Legal Compliance**: Laws like GDPR (EU), CCPA (California), and HIPAA (health data) impose strict requirement

Why This Matters for AI & Society

- **Privacy ≠ Anti-Innovation**: internet's survival no longer depends on exploitative data practices.
- **Human-Centered AI:** Regulating data sourcing is critical to **building trustworthy AI** systems.
- **Equity:** Current opt-out systems disproportionately harm vulnerable groups (e.g., those lacking time/tech literacy to navigate privacy settings)

Context, consent, and consequence

Transparency is essential when AI use impacts others, but privacy must prevail where personal autonomy or sensitive data is at risk.

The goal isn’t perfect balance it’s designing workflows where both values are intentionally protected

Data about you isn't necessarily your data

- Just because data is about you doesn't mean you own it or should have control over it

Example: Someone writing notes about you in a café doesn't make that data yours

- The claim "it's about me, therefore it's mine" is a logical fallacy

Distinguishing between enabling wrongdoing and actual wrongdoing

Collecting data that enables harmful acts isn't itself wrong

Not all data collection is surveillance

- The term "surveillance" is overused and applied too broadly
- Five levels of data collection with varying privacy violation potential (0-10 scale):
  - Level 1: Data collected but never accessed or used (0/10)
  - Level 2: Anonymized, aggregated data used impersonally (0-1/10)
  - Level 3: Someone glances at aggregated data without knowing you (1-2/10)
  - Level 4: Someone who knows you sees your data (3-8/10, depending on sensitivity)
  - Level 5: Intentional collection of sensitive data by someone who knows you (9-10/10)

How Data Privacy Is Protected

- Laws & Regulations:
  - GDPR (EU), CCPA (California), PIPEDA (Canada), LGPD (Brazil).
- Technical Measures:
  - Encryption, access controls, anonymization.
- Organizational Practices:
  - Privacy impact assessments, data audits, employee training.
- Individual Actions:
  - Using privacy tools (VPNs, ad blockers), reviewing app permissions, demanding transparency.

Current Opt-Out Systems Are Broken

Under laws like California's CCPA, users must:

- Manually request opt-outs from every individual company they've interacted with.
- Repeat this process every two years (opt-outs aren't permanent).

Result: Most people don't exercise their rights due to complexity and lack of awareness

pass data minimization and purpose limitation regulations

1. **Mandate Universal Opt-Out Signals**
   - Example: The proposed California Privacy Protection Act (CPPA) update would require all browsers to honor third-party opt-out signals (e.g., Global Privacy Control).
   - Impact: One setting blocks data collection across all websites/apps – no more per-company requests.

2. **Pass Data Minimization & Purpose Limitation Laws**
   - Data Minimization: Collect only what's strictly necessary for a stated purpose.
   - Purpose Limitation: Use data only for the purpose users consented to (e.g., not sharing health data with advertisers).

Regulate the AI Data Supply Chain

Requirements:

- Audit training data for scraped personal information.
- Prevent memorization of PII in AI outputs.
- Disclose data sources and inference risks.

Goal: Protect privacy while reducing bias and improving model reliability.

Defining “AI Use” Worth Disclosing

Threshold Test: Disclose AI use if it meets any of these criteria:

- **Impact: Alters outcomes for others** (e.g., AI-generated performance reviews, hiring recommendations).
- **Visibility: Could reasonably affect team trust or collaboration** (e.g., AI summarizing private meetings).
- **Autonomy: Reduces human control over sensitive decisions** (e.g., AI detecting "productivity" via keystrokes).

Balancing Transparency & Personal Privacy

Key Principle: Transparency without privacy is surveillance; privacy without transparency is secrecy.

Scenario	Transparency Needed?	Privacy Safeguards
AI analyzes public work data (e.g., project docs)	Low	Anonymize data; share outputs only.
AI processes team communications (e.g., Slack sentiment analysis)	High	Opt-in consent; exclude personal channels.
AI uses personal data (e.g., health insights from wearable)	Critical	Prohibit unless legally required; strict purpose limitation.

Organizational Policies, Trust & Culture

- Adopt "AI Disclosure Labels" (e.g., "AI-Assisted," "AI-Generated") for visible work.
- Create "AI Privacy Zones": Spaces/tools where AI use is banned (e.g., 1:1 meetings, HR discussions).
- Mandatory Training: Teach teams to spot AI privacy risks (e.g., "Is this tool scraping personal data?")

What can organizations do?

- Psychological Safety: Reward teams for flagging ethical concerns (e.g., "I noticed this AI tool might bias our hiring data").
- Co-Create Policies: Involve employees in drafting AI guidelines—this boosts buy-in
- Lead by Example: Managers should disclose their own AI use first.

Practical Frameworks for Disclosure

- The 3-Step Disclosure Protocol:
  - **Assess**: "**Does this AI use affect others' work, rights, or data**?"
  - Decide:
    - If YES → Disclose before use (e.g., "I'll use AI to draft this proposal—review needed").
    - If NO → No disclosure needed, but document internally.
  - Document: Log AI use in shared tools (e.g., project management software) for audit trails.
- Tools to Implement:
  - Browser Extensions: Flag AI use in real-time (e.g., "This email was drafted with AI").
  - Checklists: "Before hitting send: Did I disclose AI use? Is sensitive data excluded?"

"The future isn’t choosing between privacy and transparency. It’s designing AI that respects both."

Future Norms (Next 5 Years)

Predictions:

- Regulatory Shifts: Laws will mandate "AI disclosure" for high-stakes work (e.g., EU AI Act's workplace restrictions).
- Tech Evolution: "Privacy-First AI" tools will dominate (e.g., on-device processing, zero-data-retention).
- Cultural Change: Transparency will become default—like citing sources in research.

Prepare Now:

- Audit Your Stack: Map which AI tools access personal/team data.
- Pilot "AI Ethics Committees": Cross-functional teams to review high-risk AI use.
- Advocate for Standards: Push vendors for built-in transparency features (e.g., "Show me what data this AI used").

Practical Application

The "AI Ingredient Label" System

How it works: Treat AI-assisted work like food products—require clear "ingredients" for anything shared externally or cross-functionally.

Practical Tool: Simple tags like:

- `[AI-Generated]` (e.g., full draft by AI)
- `[AI-Assisted]` (e.g., human edited AI output)
- `[AI-Analyzed]` (e.g., data insights from AI tools)

The "Transparency-Privacy Slider" Framework

```
[High Transparency] ←————→ [High Privacy]
- Share AI use openly - Never reveal AI use
- Public documentation - Encrypted/local-only tools
- Opt-in data sharing - Zero-data-retention settings
```

How to use: **Teams "slide the marker" based on context**:

- ***High Transparency*: Client proposals, public reports.**
- ***High Privacy*: HR discussions, personal development plans**

So, what do we do?

- Build Ethics Into Process, Not Afterthoughts
- Create Diverse Review Teams
- Implement Transparency by Design
- Test for Bias Systematically
- Establish Clear Accountability
- Engage Affected Communities

Ethical Risks

Hallucinations &
Misinformation

Bias
Amplification

Operational Risks

Cost & Performance Balance

LLMOps Complexity

Do	Dont
Do anonymize data	Don’t assume outputs are factual