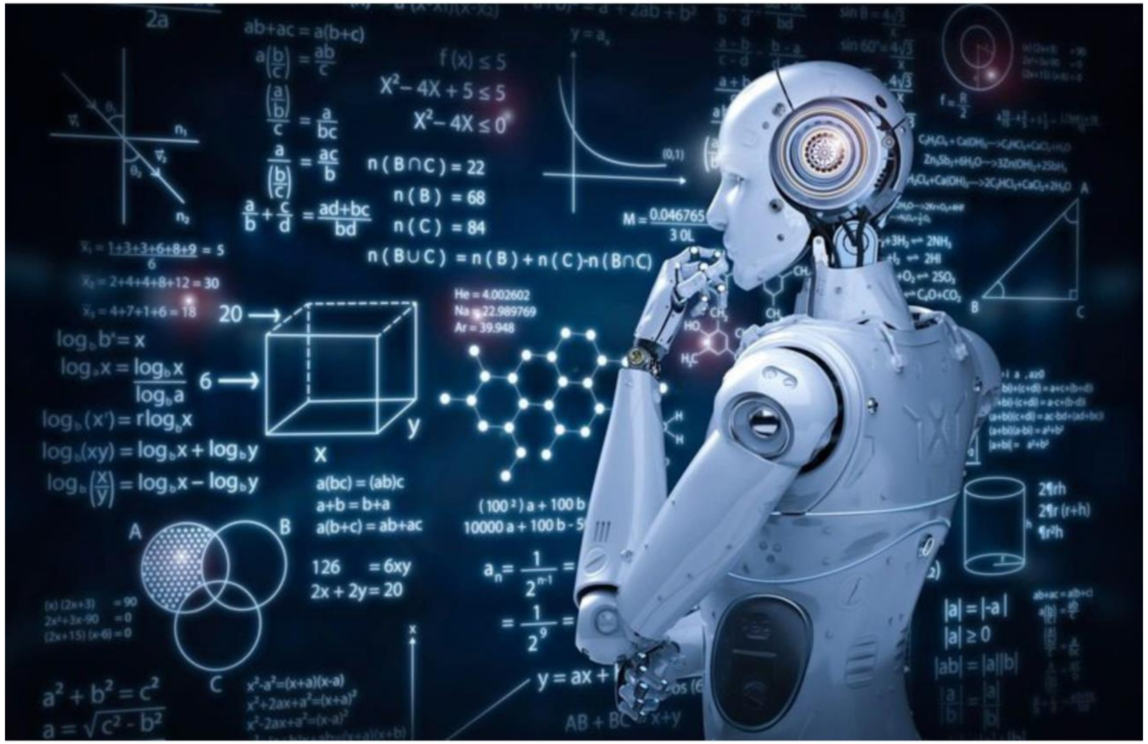
AI for Trustworthy Decision Making

Irina Mocanu

AI Decision Making

- AI to assist or fully automate choices in various domains

- AI to analyze amounts of structured and unstructured data to:
    - identify patterns,
    - predict outcomes,
    - recommend or execute decisions



AI Decision Making

- Amazon hiring AI showing **gender bias**
- Facial recognition misidentifying **people of color**
- Tesla self-driving accidents due to misinterpreted road signs

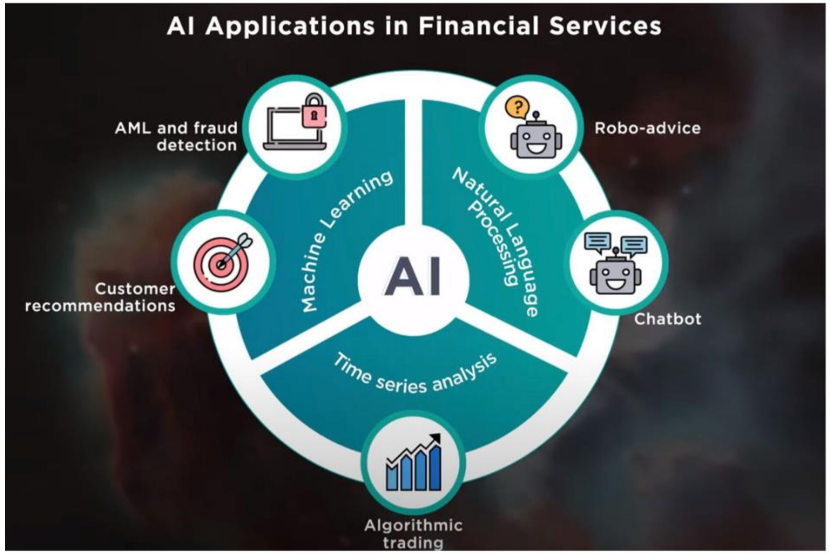
Trustworthy decisions

Healthcare

- **Diagnostic Imaging & Detection**
    - AI models assist radiologists by detecting tumors, X-rays, MRIs, and CT scans
- **Clinical Decision Support Systems**
    - Tools suggest treatment plans, potential drug interactions, or identify patients at risk of deterioration
- **Triage & Resource Allocation**
    - AI prioritizes emergency cases based on predicted severity, ensuring critical patients are seen first and guiding allocation
- **Predictive Analytics for Disease Progression**
    - Prediction of disease evolution enabling early intervention.
- **Personalized Medicine**
    - Recommend tailored therapies or predict patient response to certain drugs.

Finance & Banking

- **Credit Risk Assessment**
  - AI models analyze financial and behavioral variables beyond traditional credit scores to decide whether to approve loans or set interest rates
- **Fraud Detection & Prevention**
  - Real-time anomaly detection systems flag suspicious transactions and decide whether to block, verify, or allow them
- **Stock trading**
  - Buy/sell decisions are made based on market signals, sentiment analysis, and macroeconomic indicators
- **Customer Service**
  - AI-powered chatbots and virtual assistants decide how to respond to customer queries, route issues to human agents, or preemptively offer solutions

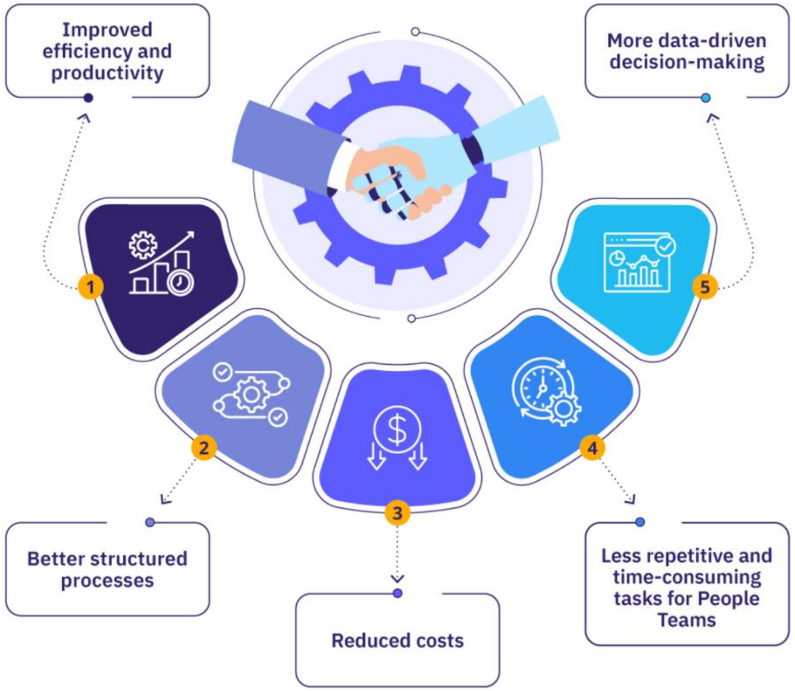


Transportation & Mobility

- **Autonomous Vehicle Navigation**
  - Self-driving platforms make continuous real-time decisions on steering, lane changes, speed adjustments based on camera, radar, and LiDAR data
- **Traffic Flow Optimization**
  - AI traffic control systems decide signal timings, reroute vehicles, and adjust lane allocations to reduce congestion and emissions
- **Public Transit Scheduling**
  - Predictive algorithms forecast passenger demand to decide optimal bus, train, or metro dispatch times, improving efficiency and reducing overcrowding.
- **Micro-Mobility & Ride-Hailing Demand Prediction**
  - Apps that forecast ride demand and decide where to position vehicles or scooters in real time to maximize availability and revenue.

Human Resources & Recruitment

- **Automated Resume Screening**
  - Platforms parse resumes, evaluate keywords, and score candidates to decide who advances to interviews
- **Job Matching & Internal Mobility**
  - AI-driven systems suggest positions to applicants or recommend internal role changes for current employees based on skill gaps and career paths
- **Workforce Planning**
  - Predictive AI tools forecast future talent needs and decide optimal hiring timelines based on business growth and market trends.


Improved efficiency and productivity
More data-driven decision-making
1
5
2
4
3
Better structured processes
Less repetitive and time-consuming tasks for People Teams
Reduced costs

https://www.aihr.com/blog/ai-in-hr/

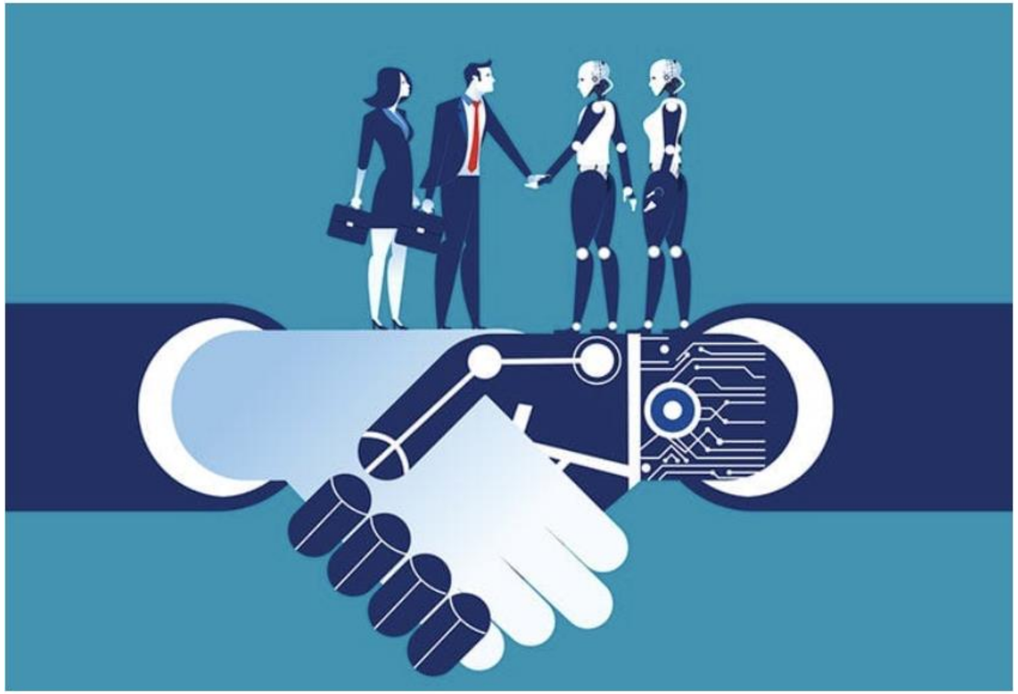
Why AI Trust Matters

- AI is increasingly supporting decisions in different domains: healthcare, finance, law, and governance.
- **With trust**: people will accept and follow AI-assisted decisions.
- **Without trust**: low adoption, high resistance, and risk of societal backlash.

- AI’s decisions and actions must be understood, predicted, and controlled.

Why AI Trust Matters

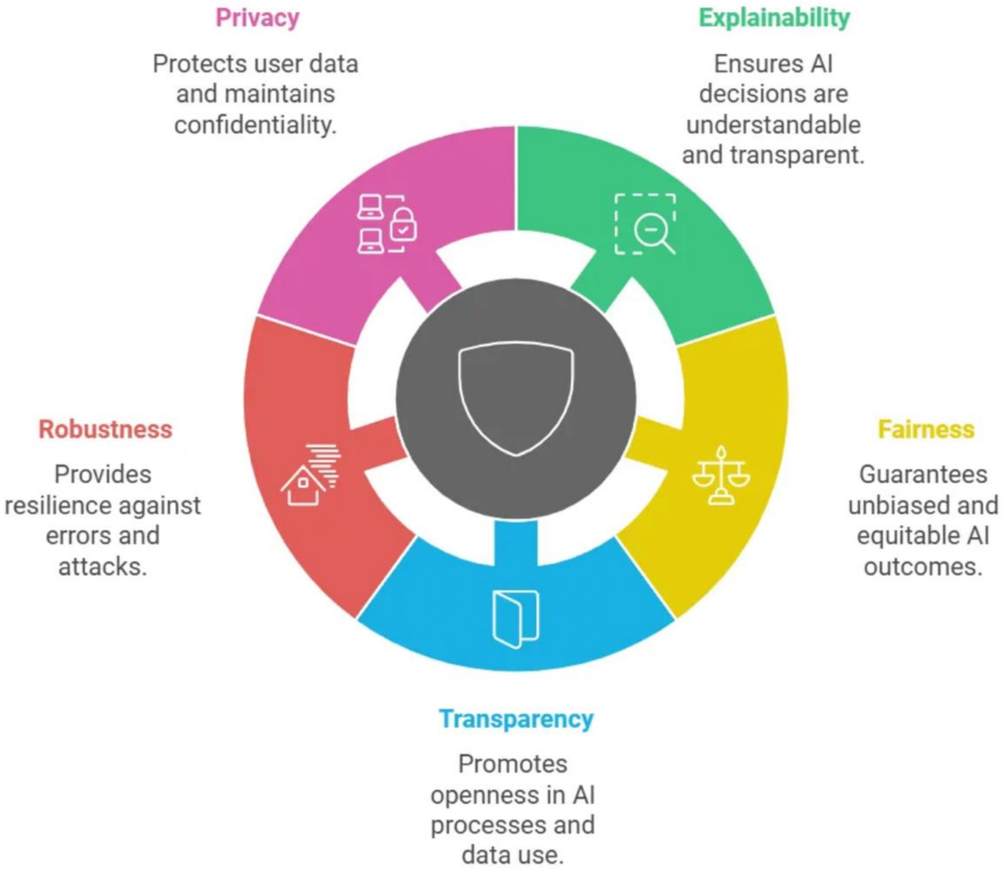
- Erosion of customer trust and confidence
  - An AI chatbot that is unable to comprehend and accurately respond to customer queries - can lead to frustration, dissatisfaction, and a potential loss of loyal customers.
  - Unreliable AI solutions can compromise the privacy and security of sensitive customer data
- AI systems that prioritize transparency, reliability, and ethical considerations.



Principles of Trustworthy AI

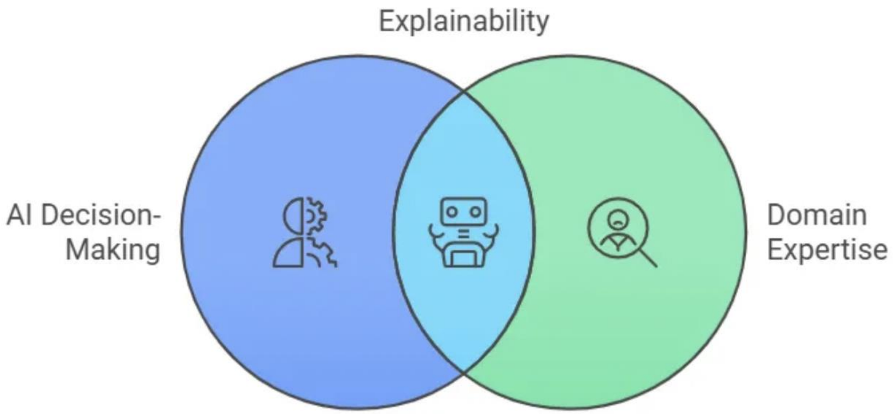
IBM defined five principles: pillars of trustworthy

- Explainability
- Fairness
- Transparency
- Robustness
- Privacy


Privacy
Protects user data and maintains confidentiality.
Explainability
Ensures AI decisions are understandable and transparent.
Robustness
Provides resilience against errors and attacks.
Fairness
Guarantees unbiased and equitable AI outcomes.
Transparency
Promotes openness in AI processes and data use.

https://www.youtube.com/watch?v=nB_EjxoP-6w
https://medium.com/@tahirbalarabe2/five-principles-of-trustworthy-generative-ai-artificial-intelligence-models-67b58402d40a

Principles of Trustworthy AI



Explainability

- AI can justify its decisions in a way that makes sense
- Chatbot: I have red itchy eyes, a runny nose, and I am sneezing
  - A doctor - signs of an allergy.
  - A trustworthy AI – the same response.
- Chatbot: you have a broken leg: that's not explainable.

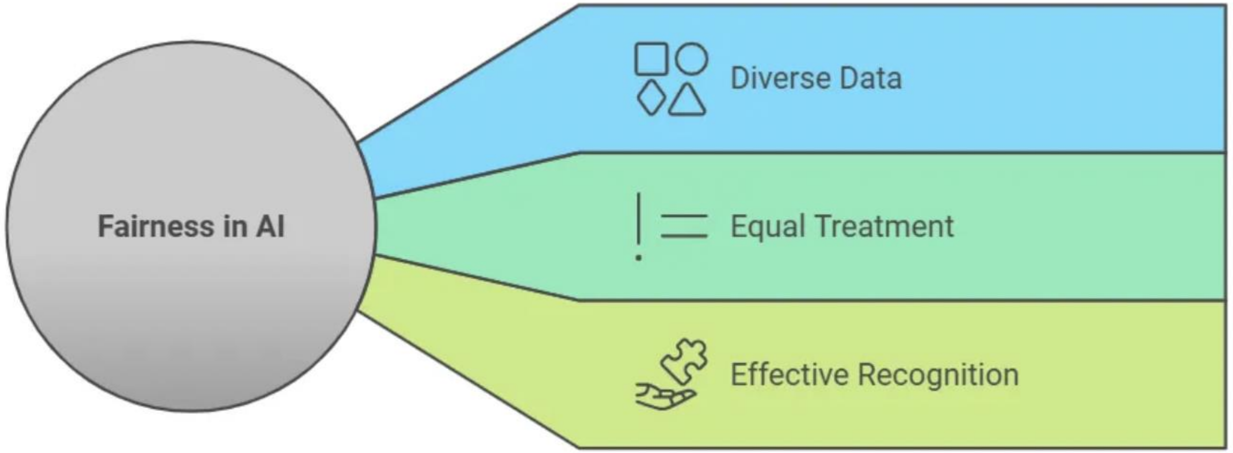
- A domain expert (a doctor) — should be able to look at the AI's reasoning and say, "Yes, I see how it got there." They shouldn't need to be an AI expert.

https://www.youtube.com/watch?v=nB_EjxoP-6w
https://medium.com/@tahirbalarabe2/five-principles-of-trustworthy-generative-ai-artificial-intelligence-models-67b58402d40a

Principles of Trustworthy AI

Fairness

- AI treats everyone equally – AI needs diverse data
- Train an AI to recognize shapes: only pictures of squares are used.
- Show a circle, it’s likely to fail: it doesn’t know what to do because it hasn’t seen enough diverse examples.

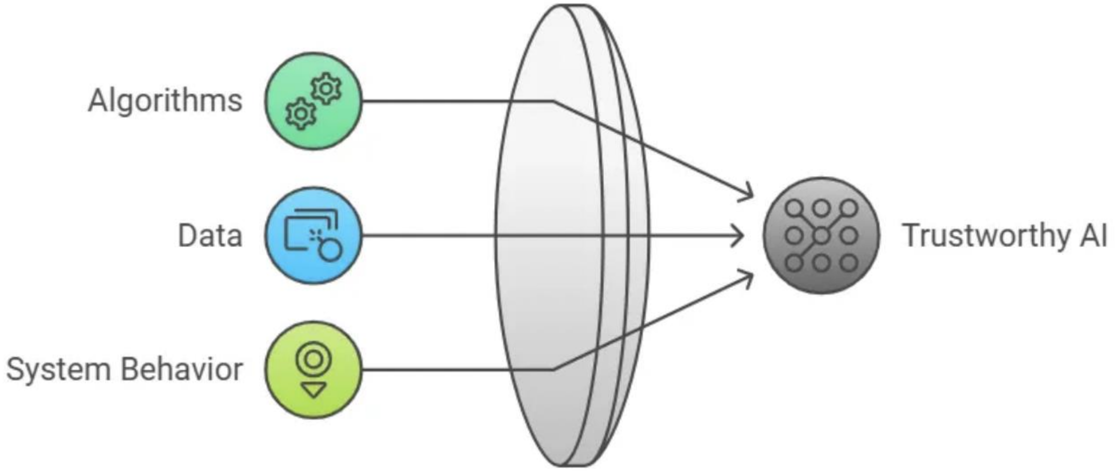


https://www.youtube.com/watch?v=nB_EjxoP-6w
https://medium.com/@tahirbalarabe2/five-principles-of-trustworthy-generative-ai-artificial-intelligence-models-67b58402d40a

Principles of Trustworthy AI

Transparency

- A system that tells you, “Trust me.”
- To trust it: to see inside:
  - What algorithms is it using?
  - What data was it trained on?
- A transparent AI: can check its work
  - where it came from,
  - how it was built,
  - why it behaves the way it does.

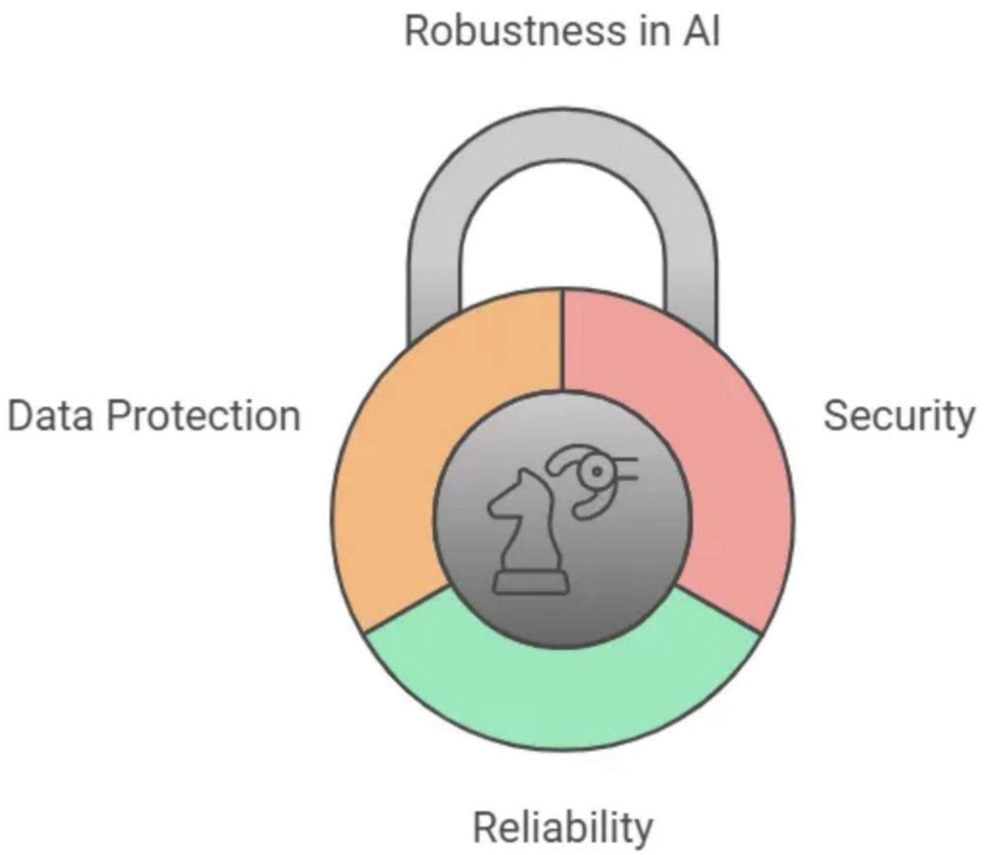


https://www.youtube.com/watch?v=nB_EjxoP-6w
https://medium.com/@tahirbalarabe2/five-principles-of-trustworthy-generative-ai-artificial-intelligence-models-67b58402d40a

Principles of Trustworthy AI

Robustness

- Staying secure and reliable.
- A robust AI can fend off attempts to steal or poison its data.
- If an AI can be manipulated, its outputs can’t be trusted.

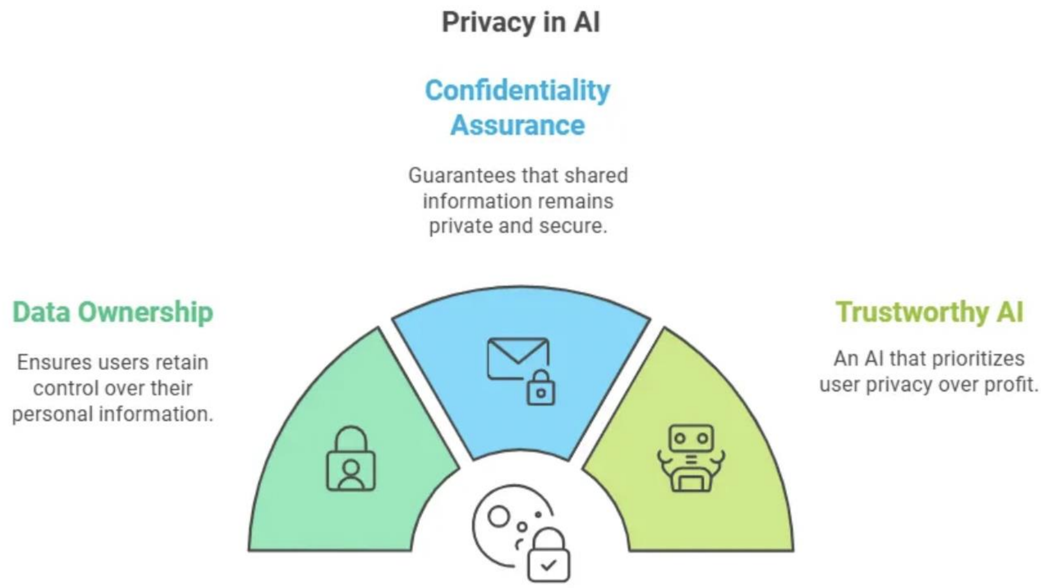


https://www.youtube.com/watch?v=nB_EjxoP-6w
https://medium.com/@tahirbalarabe2/five-principles-of-trustworthy-generative-ai-artificial-intelligence-models-67b58402d40a

Principles of Trustworthy AI

Privacy

- Your data stays yours
- If you share something with a chatbot, you shouldn't worry that it will be leaked or sold.
- A trustworthy AI keeps information private.
- Data isn't its business model.



Privacy in AI
Confidentiality Assurance
Guarantees that shared information remains private and secure.
Data Ownership
Ensures users retain control over their personal information.
Trustworthy AI
An AI that prioritizes user privacy over profit.

https://www.youtube.com/watch?v=nB_EjxoP-6w
https://medium.com/@tahirbalarabe2/five-principles-of-trustworthy-generative-ai-artificial-intelligence-models-67b58402d40a

Trust & Trustworthiness

- **Trust**
- Confidence that an AI system will act as expected.
- Subjective based on perception, reputation, and user experience.
- It can exist even if it is not justified.
- **Example:**
  People may trust a navigation app blindly, even if it sometimes gives wrong directions.

Trustworthiness

- The actual quality of being reliable, fair, transparent, ethical, and accountable.
- Can be measured, verified, and evaluated with evidence.
- Independent of whether people perceive it.
- **Example:**
  A medical AI system that is demonstrably unbiased, accurate, and privacy-preserving is trustworthy — even if patients don’t yet trust it.

Trust & Trustworthiness

Healthcare

- **Trust:** A patient may trust a symptom-checker app because it has a nice interface and feels reliable
- **Trustworthiness:** AI might be **trustworthy** (tested, validated, unbiased), but patients may still hesitate to trust it due to fear of “machines replacing doctors.”

Trust & Trustworthiness

Finance

- **Trust:** A customer might trust a flashy fintech app’s loan approval system because it seems quick and modern.
- **Trustworthiness:** a credit-scoring model built with fairness constraints could be highly trustworthy, but customers may remain skeptical if they don’t understand the decision process.

Trust & Trustworthiness

Self-Driving Cars

- **Trust:** A driver may over-trust autopilot and stop paying attention, assuming the AI can handle all conditions.
- **Trustworthiness:** The car might not be safe in bad weather or unusual traffic scenarios, showing it isn’t fully trustworthy yet.

Trust & Trustworthiness

- **Trust = belief** (subjective, user-side).
- **Trustworthiness = reality** (objective, system-side).

- People may **over-trust** untrustworthy systems (e.g., black-box AI with hidden biases).
- People may **under-trust** trustworthy systems (e.g., safe medical AI rejected because it's "too opaque").

- **Trust (belief)** can exist **without trustworthiness (reality):** dangerous over-reliance.
- **Trustworthiness (reality)** can exist **without trust (belief)**: under-use of valuable AI.
- Goal = align both.

Factors Affecting Trust

- **Technical trustworthiness:** performance, robustness, privacy, fairness
- **Human factors:** explainability, usability: bridge the gap between AI and users
- **Social and institutional context** determines whether people accept and rely on the system.

Factors Affecting Trust

Technical Factors

- **Performance & Reliability**
  - Robustness under noisy data, stability across time
  - Example: AI weather prediction system consistently forecasts rainfall with >90% accuracy.
- **Transparency & Explainability**
  - Use of interpretable & explainable models, layperson-friendly explanations
  - Example: A credit-scoring AI explains which financial behaviors (late payments, income stability) most influenced the decision.
- **Fairness & Bias Mitigation**
  - Bias in datasets and algorithms
  - Example: An AI hiring tool ensures equal opportunity across genders and ethnic groups.
- **Security & Privacy**
  - Data protection mechanisms: privacy, attacks
  - Example: training models without sharing health data
- **Accountability Mechanisms**
  - Model versioning, clear responsibility for system errors
  - Example: A self-driving car company logs all AI decisions so that failures can be investigated and responsibility assigned.

Factors Affecting Trust

Human-Centric Factors

- **Human Oversight & Control**
  - Human-in-the-loop for critical decisions
  - Ability to override or contest AI decisions
  - Example: A doctor reviews AI-generated cancer diagnoses before finalizing treatment
- **Usability & User Experience**
  - Intuitive interfaces that support decision-making
  - Interactive feedback mechanisms to improve system adoption
  - Example: Farmers use a mobile AI app that provides pest control advice in simple visuals and local language.
- **Perceived Transparency**
  - Even if explainability exists, users must understand it
  - Example: A banking chatbot explains fees in plain, understandable terms
- **Trust Calibration**
  - Preventing over-trust / under-trust (blind reliance on AI / ignoring useful AI)
  - Training users to interpret AI outputs correctly
  - Example: Airline pilots are trained to know when to trust autopilot and when to intervene manually

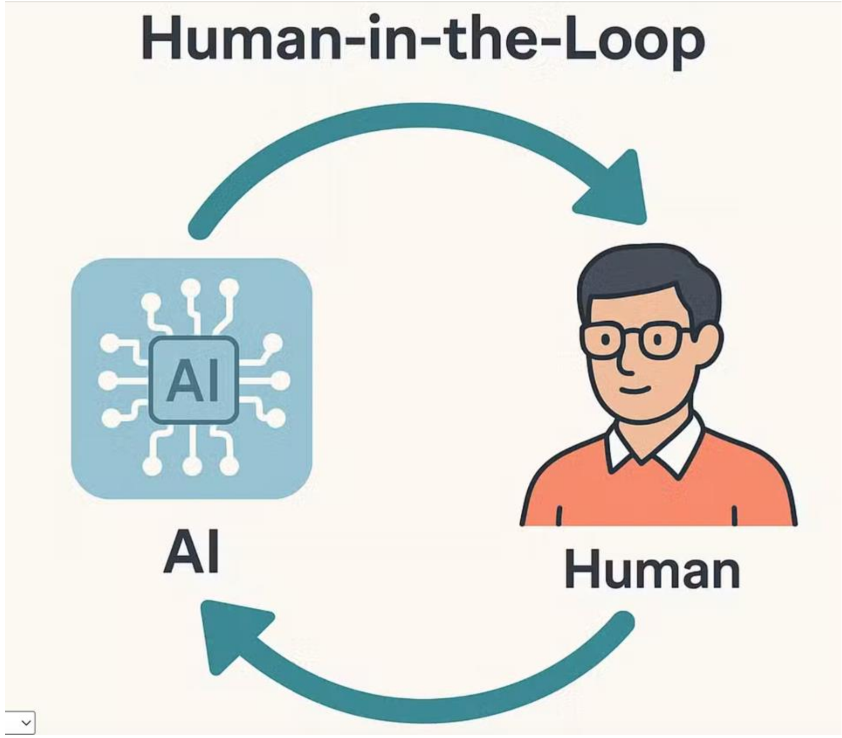
Factors Affecting Trust

Social, Ethical & Institutional Factors

- **Ethical Alignment**
    - Respect for human rights, autonomy, and dignity, and cultural norm
    - Example: An AI recruitment tool excludes sensitive attributes (e.g., race, religion) from decision-making
- **Reputation & Source Credibility**
    - Trust in the organization developing/deploying AI
    - Certification by trusted institutions
    - Example: AI medical devices approved by the FDA
- **Legal & Regulatory Compliance**
    - Adherence to AI-specific laws (GDPR), audits and certifications
    - Example: European AI chatbots comply with GDPR, protecting user data rights.
- **Social Acceptance & Cultural Context**
    - Community involvement in design and governance
    - Example: In Japan, AI caregiving robots are widely accepted in elderly care, while other societies may resist them.
- **Risk Communication & Public Engagement**
    - Involving stakeholders in AI design and evaluation
    - Example: A government explains the risks and benefits of AI traffic surveillance in public consultations before deployment

Decision-Making with AI

- **Fully Automated**: AI decides without human involvement.
- **Human-in-the-Loop**: AI assists, but humans make final decisions.
- **Human-on-the-Loop**: AI makes decisions, humans monitor & can intervene.



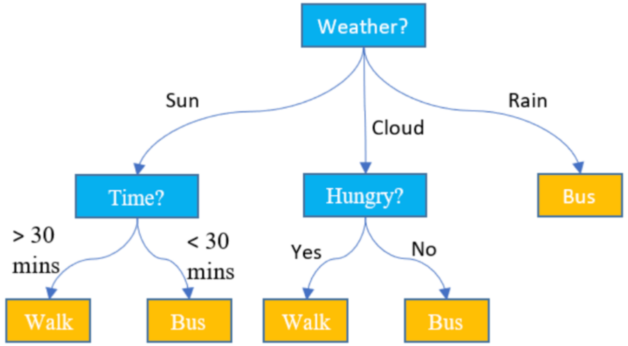
Risks of Low Trust

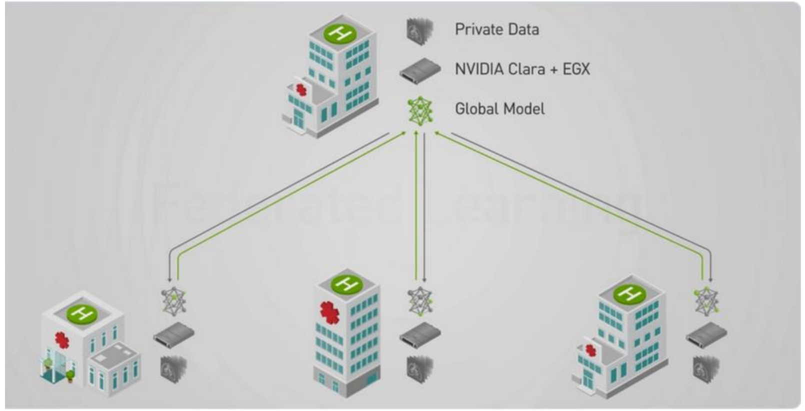
- Rejection of beneficial AI tools
- Misuse due to over-trust (“automation bias”)
- Legal and ethical disputes over AI decisions
- Damage to institutional credibility

AI Models Representative of Trust in Decision-Making

- **Models where trust is essential** - important decisions.

- **Models designed to enhance trustworthiness** - through interpretability, fairness, or privacy.





AI Models Representative of Trust in Decision-Making

Rule-Based Expert Systems

- **Example: Law & Taxation:**
  - *Input: income, no of children, luxury assets*
  - *Rule 1: IF income < €10,000 AND dependent children = yes, THEN tax exemption = true.*
  - *Rule 2: IF income > €50,000 AND luxury assets = yes, THEN extra tax applies*.
  - Default: If no rules match: standard tax rules apply (no exemption, no extra luxury tax flagged)

- A lawyer or auditor can directly verify the logic.
- Even if it's not as flexible as ML, it guarantees accountability.

- Encode human expertise as a set of "if–then" rules.
- The rules are explicitly written, not learned from data (unlike ML models).

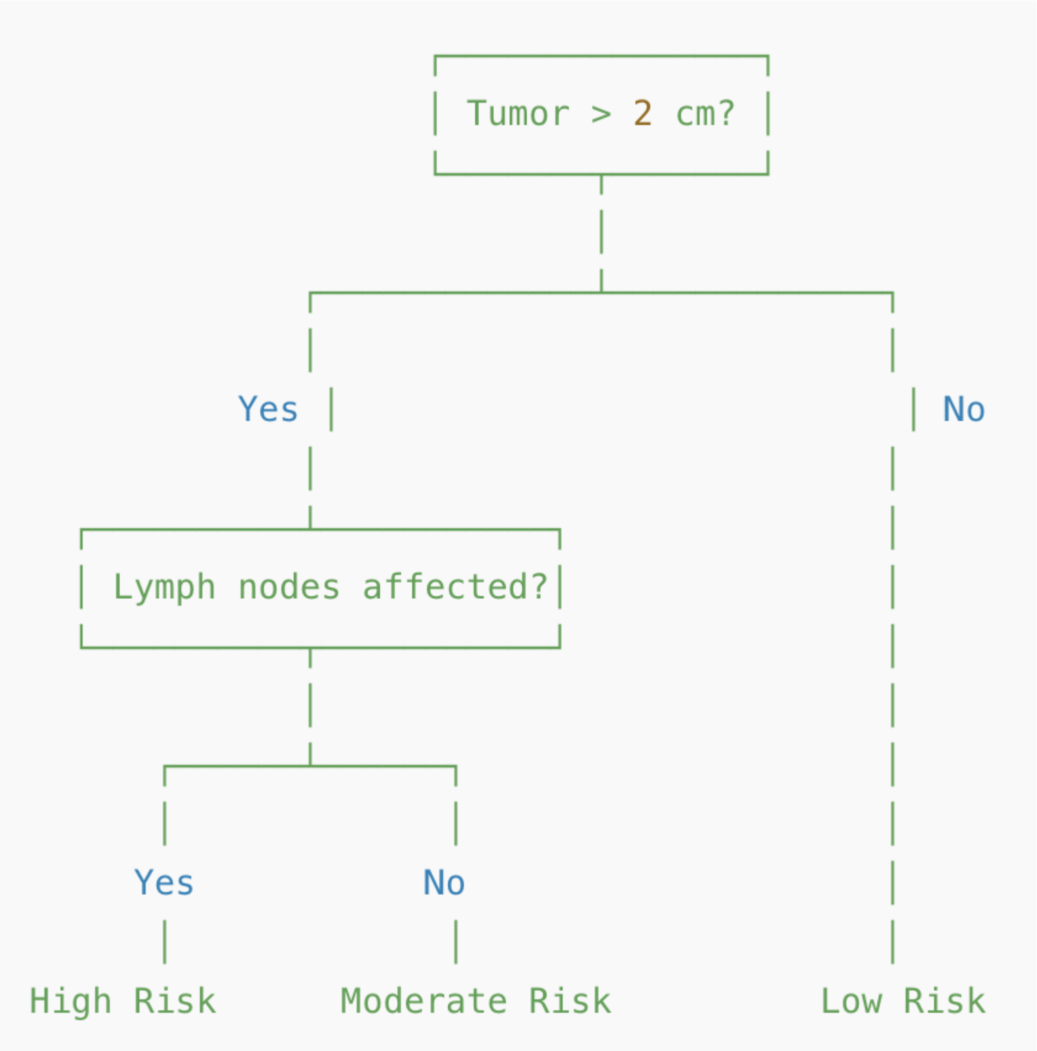
- Decisions are *transparent* by design: each outcome is directly traceable to a rule.
- They align with legal or regulatory frameworks where rules are predefined.

AI Models Representative of Trust in Decision-Making

Decision Trees

- **Example:** breast cancer diagnosis:

- Each path is human-readable: easy to explain to a doctor
- Doctors can verify this logic against medical guidelines: the AI is more trustworthy

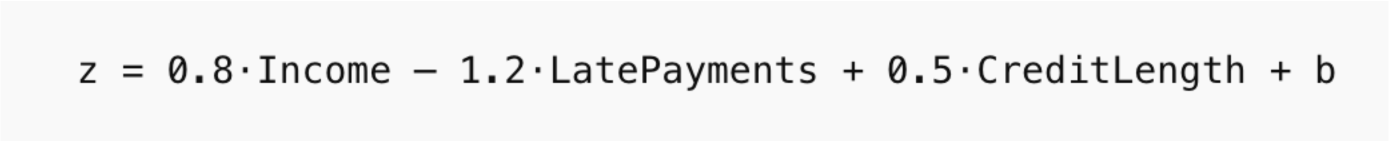


AI Models Representative of Trust in Decision-Making

Logistic Regression

- **Example: Finance:** A bank uses logistic regression for loan approval:
  - Inputs:
    - Income
    - Late payments
    - Credit history length
  - Output:
    - Calculates the probability of an outcome using a weighted sum of input features. - a probability between 0 and 1 (e.g., 80% chance of loan approval).

    - Income: weight = +0.8 → higher income strongly increases approval odds.
    - Late payments: weight = -1.2 → past defaults strongly decrease odds.
    - Credit history length: weight = +0.5 → longer history slightly increases odds.



```
z = 0.8·Income − 1.2·LatePayments + 0.5·CreditLength + b
```

$\xrightarrow{1 / (1 + e\text{^}-z)}$ Loan Approval Probability

- Coefficients are interpretable: they show *how much each factor contributes* to the decision.

- Decisions can be justified to customers: “you were denied because of repayment history, not gender/ethnicity”

AI Models Representative of Trust in Decision-Making

1. Interpretable / Transparent Models

- **Rule-Based Expert Systems:** human-readable rules
- **Decision Trees:** follow step-by-step
- **Logistic Regression:** understandable weights for features

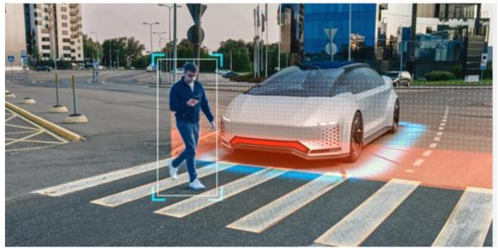
- Easier for humans to trust or contest decisions.

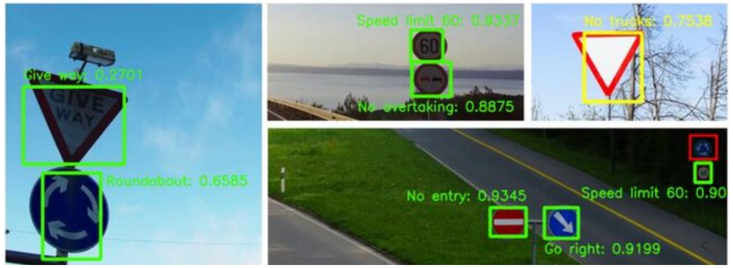
- They offer **clarity and justification**.

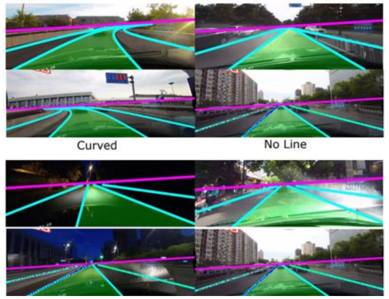
AI Models Representative of Trust in Decision-Making

Deep Neural Networks

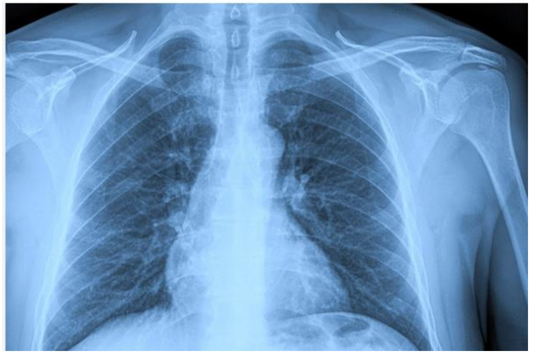
- **Examples:**
  - **Self-driving cars**

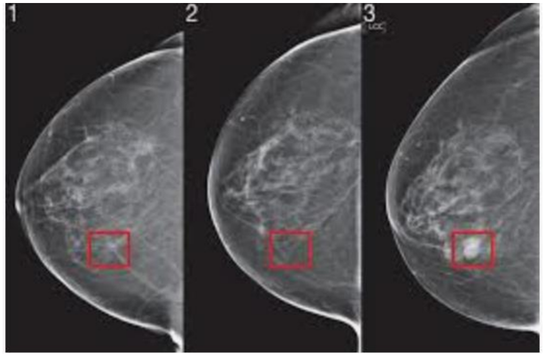






  - **Medical imaging**





- The reasoning inside hidden layers is not directly interpretable: “black box.”
- Hard to answer *why* the model made a decision.

AI Models Representative of Trust in Decision-Making

Ensemble Models (Random Forest, XGBoost)

- Combine predictions from many “weak learners” to improve accuracy

- **Examples:**
  - **Fraud detection:** identify unusual transactions
    - Customers denied a purchase rarely get an understandable explanation.
  - **Credit scoring:** predict a credit
    - A decision may depend on subtle interactions between features, not obvious to users.

  - Individual trees are interpretable, but hundreds or thousands of them are not.

AI Models Representative of Trust in Decision-Making

- **Transformers (NLP):** capture relationships between words in a sequence
- **Examples:**
  - **Healthcare: clinical note summarization:** summarize patient history for doctors.
    - *Trust challenge:* Omitted or wrongly emphasized details could affect treatment.
  - **Customer service chatbots (banking, airlines, telecoms):** Handle queries 24/7.
    - *Trust challenge:* Incorrect explanations about fees or policies could erode customer trust.

  - They can generate *plausible but false* answers (“hallucinations”).
  - Difficult to verify the reliability of reasoning.

AI Models Representative of Trust in Decision-Making

2. Black-Box but High-Performance Models

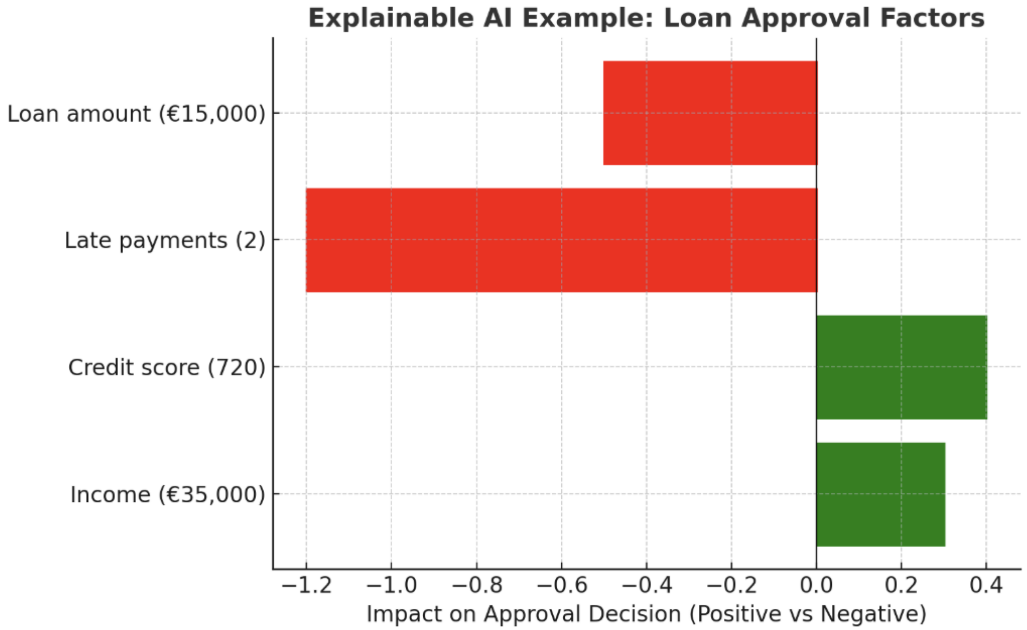
- **Deep Neural Networks**
- **Transformer Models**
- **Ensemble Models**

- They perform well but raise **trust challenges** due to lack of transparency

AI Models Representative of Trust in Decision-Making

- Explainable AI: For a customer:
  - Income = €35,000
  - Credit score = 720
  - Past late payments = 2
  - Loan amount = €15,000
- The model predicts: **Loan Rejected**

Feature	**Impact on Decision**
Income (€35,000)	+0.3 (positive)
Credit score (720)	+0.4 (positive)
Late payments (2)	-1.2 (negative)
Loan amount (€15,000)	-0.5 (negative)



AI Models Representative of Trust in Decision-Making

Bias

- Unfair decisions (e.g. biased hiring tools)
- Reduces accuracy & generalization: model fails in real-world scenarios
- Create legal & ethical risks (discrimination lawsuits, loss of trust).

- **Dataset bias:** training data underrepresents minorities.
- **Algorithmic bias:** model unfairly favors one group.
- **Feedback loops:** predictive policing targeting areas more heavily policed.

AI Models Representative of Trust in Decision-Making

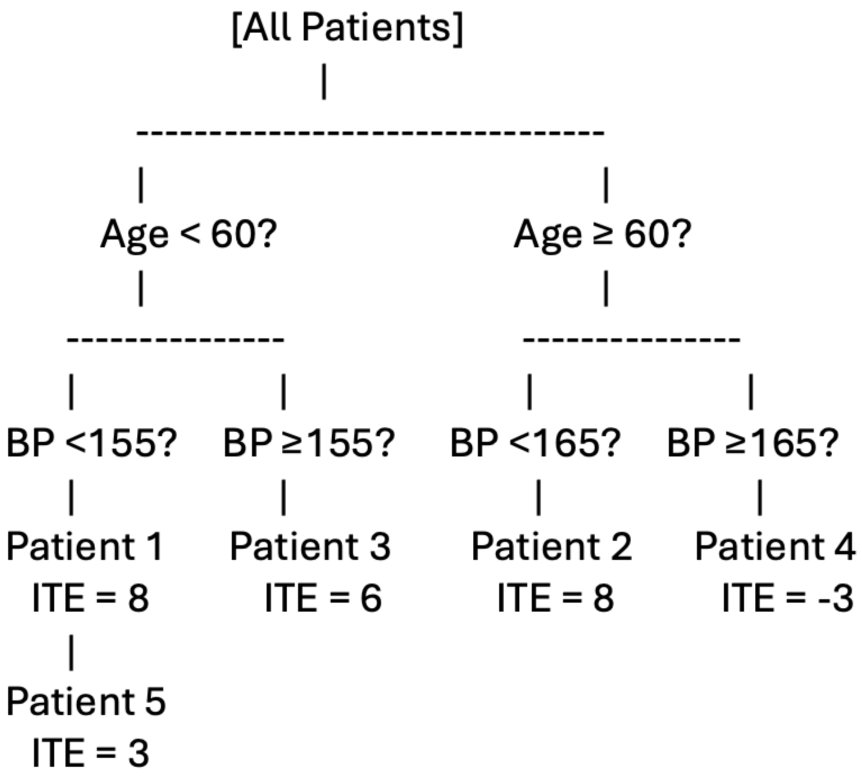
- **Selection Bias**
  - Data collected doesn’t represent the target population
  - Example: A face recognition system trained mostly on lighter-skinned faces may perform poorly on darker-skinned faces.
- **Label Bias**
  - Labels are assigned or annotated with inconsistencies
  - Example: Different doctors labeling X-ray scans in a different way
- **Exclusion Bias**
  - Important groups or features are not considered during data preparation
  - Example: A health study excludes older patients: the model can’t make accurate predictions for seniors

AI Models Representative of Trust in Decision-Making

- **Fair ML Models (with bias mitigation)**
  - **resampling, or fairness constraints** are applied to reduce bias in data or model outputs.
  - ensure predictions do not unfairly disadvantage certain groups.
  - corrects for bias in data/models
- **Examples:**
  - **Hiring AI:** Tools that match candidates to jobs, adjusted so gender or ethnicity does not bias recommendations.
  - **University admissions:** Reweights applications so that socioeconomic background does not unfairly lower acceptance chances.
  - **Bank lending models:** Adjust loan approval systems to remove systemic racial bias from historical data.

AI Models Representative of Trust in Decision-Making

- **Causal Models**
  - decisions are based on causal evidence, not on correlations
  - Received Drug
  - Age, Blood Pressure, Medication, Observed Outcome (BP Reduction)



[All Patients]

Age < 60?

Age ≥ 60?

BP <155?

BP ≥155?

BP <165?

BP ≥165?

Patient 1 ITE = 8

Patient 3 ITE = 6

Patient 2 ITE = 8

Patient 4 ITE = -3

Patient 5 ITE = 3

AI Models Representative of Trust in Decision-Making

- **Causal Models**
  - decisions are based on causal evidence, not on correlations
- **Examples:**
  - **Healthcare treatment analysis:** Determining if a new drug *causes* better recovery rates, controlling for patient age and health.
  - **Education policy:** Identifying whether smaller class sizes *cause* improved student performance (vs. just correlated).

AI Models Representative of Trust in Decision-Making

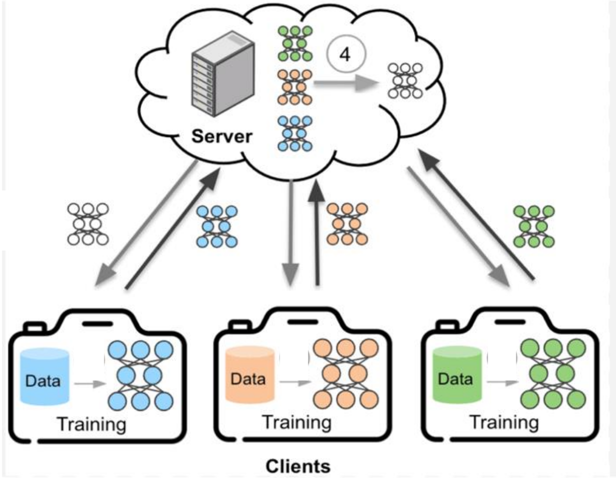
3. Fairness- and Accountability-Oriented Models

- **Fair ML Models** (with bias-mitigation constraints): Adjust predictions to ensure equal opportunity
- **Causal Models:** Used to infer cause-effect relationships (important in law and healthcare where accountability matters).

- Fairness and accountability are embeded

AI Models Representative of Trust in Decision-Making

- **Federated Learning**
  - AI models are trained across **decentralized devices/servers**, keeping raw data local.
  - Only model updates (not raw data) are shared and aggregated.



Examples:

Healthcare: Train AI models with data from multiple hospitals without sharing patient records

Mobile devices: Keyboard suggestions trained on local typing data without sending messages to servers

Banking consortia: Multiple banks collaborate on fraud detection models without exposing customer data

Sensitive data (e.g., patient records, financial transactions) never leaves its source
Builds trust by reducing risk of data breaches and respecting privacy laws (GDPR)

AI Models Representative of Trust in Decision-Making

Differential Privacy

  - Adds noise to data or model outputs so individuals cannot be identified

- **Examples:**
  - **Tech companies (Apple, Microsoft, Google):** Collect usage statistics (e.g., which emoji is most popular) without identifying individual users
  - **Recommendation systems (Netflix, Spotify):** Analyze user preferences without revealing exact viewing or listening habits

- Ensures personal information stays confidential even if datasets are released
- Supports compliance with privacy regulations

AI Models Representative of Trust in Decision-Making

4. Privacy-Preserving Models

- **Federated Learning:** AI learns across distributed data without centralizing sensitive information (healthcare, finance)
- **Differential Privacy Models:** Add noise to protect individual data while allowing insights (recommendation systems)

- They preserve **trust through privacy** — critical in sensitive decision domains.

- **Federated Learning:** Trust through decentralization (no central data exposure)
- **Differential Privacy:** Trust through mathematical privacy guarantees

AI Models Representative of Trust in Decision-Making

Reinforcement Learning with Human Feedback

- AI agents learn not just from rewards in the environment, but also from human preferences or corrections

- **Examples:**
  - **ChatGPT & LLMs (OpenAI, Google DeepMind):** responses follow human ethical guidelines
  - **Robotics (Boston Dynamics experiments):** Robots learn tasks (e.g., grasping fragile objects) with human-provided corrections instead of trial-and-error alone

AI Models Representative of Trust in Decision-Making

- **Interactive Machine Learning**
- The user iteratively provides feedback, corrections, or labeling during training and model refinement
- **Examples:**
  - **Medical diagnosis assistants:** Radiologists correct AI’s mislabeling of X-rays; the AI adapts to the expert’s style over time
  - **Personalized recommendation systems (Spotify, Netflix, Amazon):** User feedback iteratively shapes the recommendation model

AI Models Representative of Trust in Decision-Making

- **Hybrid Decision Systems (AI proposes, human confirms)**
- AI makes a prediction or recommendation, but the final authority remains with a human.

- **Examples:**
  - **Air traffic control**: AI suggests route adjustments for safety and efficiency; controllers validate them before action
  - **Clinical decision support (Mayo Clinic trials):** AI proposes cancer treatment options; oncologists approve or modify recommendations.

AI Models Representative of Trust in Decision-Making

5. Human-in-the-Loop Models

- **Reinforcement Learning with Human Feedback:**
  - Used in AI alignment (e.g., aligning large language models with ethical guidelines)
  - Aligns AI with human values
- **Interactive ML Systems:**
  - Users iteratively guide model updates (e.g., medical diagnosis assistants)
  - Human users refine and shape AI behavior in real time
- **Hybrid Decision Systems:**
  - AI proposes, human confirms

- Balance **automation with oversight**, crucial for calibrated trust.

Table of Contents

- **Introduction**
- **AI Methods Relevant to Trustworthy Decision-Making**
  - Machine Learning (Supervised, Unsupervised)
  - Privacy-Preserving AI
    - Federated Learning, Differential Privacy, Secure Multi-party Computation
  - Explainable AI Methods
  - Fairness-Aware Machine Learning
    - Bias detection and mitigation techniques
  - Human-in-the-Loop AI and Hybrid Systems
- **Evaluating Trustworthiness**
  - Metrics for Reliability, Accuracy, and Robustness
  - Fairness and Bias Evaluation Metrics
  - Explainability and Transparency Measures
  - Accountability and Governance Models
- **Challenges and Limitations**
- **Applications: healthcare**

Grades

- Project: 60 points
    - A set of datasets
    - Projects made in teams (3-4 students)
        - Starting from a set of datasets (medicine and human action recognition)
        - Apply federated learning
        - Add: explainability and bias data
        - Each student must present his/her contribution
        - 2-3 physical presentation
- Exam: 40 points
- Bonus: 10 points (in-class activity)