RAI Crisp Notes

Ch1:

What's Responsible AI?

Responsible AI (RAI) means **humans** need to make sure AI is **ethical**, **fair**, and **safe**. Since AI is a tool, **people** (not machines) are responsible for its actions. If AI makes a mistake (like a **self-driving car** crashing), it's the **creators** or **users** who should be held accountable.

Key Questions:

- Can AI be **biased**?
- Will it take **jobs**?
- Who's responsible if AI makes a bad decision?

AI can't be legally responsible like humans, so we need **strong governance** to manage these issues.

What is AI?

AI is like a **smart robot brain** that processes data to make decisions. It learns through patterns (machine learning) and helps in various tasks, but it's different from human intelligence. Think of it like following a **recipe**: good ingredients (data) lead to better results.

Building Responsible AI

To build AI responsibly:

- Use **high-quality, diverse data**.
- Involve a **diverse team** of developers.
- **Educate** society about AI's impact.

Benefits & Risks of AI

Benefits: AI can improve **healthcare**, **safety**, and **productivity**, making faster and more accurate decisions.
Risks: AI raises concerns about **job loss**, **privacy**, and even **superintelligent robots**. But, at the end of the day, **humans** are in control.

Ethics in AI

There are three parts to AI ethics:

1. **Ethics in Design**: AI should consider **societal impact**.
2. **Ethics by Design**: Teach AI to follow **moral values**.
3. **Ethics for Designers**: Set rules for the **people building AI**.

Ch2:

What is AI?

- **AI**: Machines that perform tasks requiring intelligence (e.g., solving problems, answering questions).
- **Example**: Siri, Google Assistant.
- **Turing Test**: If you can’t tell if it's a machine or a human, it's intelligent.

AI Agents (Smart Helpers)

- **Reactivity**: Respond to environment (e.g., Roomba senses dirt).
- **Proactivity**: Takes action on its own (e.g., self-driving cars).
- **Sociability**: Interacts with humans/other machines (e.g., robots in factories).

History of AI

- Influenced by **Computer Science, Philosophy, Psychology**.
- AI should be **responsible**—safe for society.
- Two views:
  - **Engineering**: Solve real-world problems (e.g., healthcare).
  - **Scientific**: Understand how intelligence works.

Types of AI Systems (Russell and Norvig)

- **Think like humans** (Neural Networks).
- **Act like humans** (Humanoid robots).
- **Think rationally** (Logical problem-solving).
- **Act rationally** (Decision-making).

Philosophy of AI

- **Symbolic AI**: Based on rules (e.g., chess AI).
- **Sub-symbolic AI**: Learns from experience (e.g., neural networks).
- **Ethics**: AI in healthcare or law should be ethical.
- **Superintelligence**: AI surpassing human intelligence (still far off).

Autonomy in AI

- **Task Autonomy**: AI adjusts behavior to complete tasks (e.g., robot vacuum).
- **Goal Autonomy**: AI sets its own goals (advanced).
- **Social Autonomy**: AI works with others (multi-agent systems).

Examples

- **Roomba**: Reactive.
- **Siri**: Reactive & sociable.
- **Tesla (Self-driving)**: Task autonomy.
- **Chess AI**: Smart but limited to one task.

Ch3:

1. **Ethical Reasoning**:
    - AI must identify ethical problems, assess consequences, and justify decisions using ethical frameworks.
    - Example: A robot must decide whom to save—someone drowning or someone trapped in a fire.
2. **AI as a Moral Agent**:
    - AI should respect human values and uphold human rights.
    - Example: An AI used for hiring should avoid biases and treat candidates fairly.
3. **Ethical Theories**:
    - **Consequentialism**: Focus on outcomes—choose the action that benefits the most people.
    - **Deontology**: Follow strict rules or duties, even in complex scenarios.
    - **Virtue Ethics**: AI should act in line with human virtues like kindness and fairness.
4. **Challenges in Ethical AI**:
    - **Identifying Values**: AI must understand and balance conflicting values (like health vs. wealth).
    - **Making Decisions**: Real-time ethical decisions are tough, especially when multiple principles conflict.

Formal Definitions and Technical Jargon

1. **Ethical Reasoning**:
   The process by which AI systems evaluate various scenarios from a moral perspective, weighing the consequences of different actions and aligning their decisions with ethical frameworks.
2. **Moral Agent**:
   An entity (in this case, AI) expected to behave in a morally responsible manner, similar to how humans follow ethical norms and respect values such as fairness, equality, and justice.
3. **Consequentialism (Utilitarianism)**:
   A theory of ethics that bases the morality of an action on its outcomes or consequences. For AI, this means choosing the action that results in the greatest good for the greatest number of people.
4. **Deontology**:
   An ethical framework that emphasizes following set moral rules or duties regardless of the consequences. In AI, this would involve strict adherence to ethical principles (e.g., honesty, no harm).
5. **Virtue Ethics**:
   A theory that focuses on the moral character of the agent rather than specific actions. In AI, this implies that systems should be designed to mimic virtues like kindness, fairness, and empathy.
6. **Values Identification**:
   The ability of AI to recognize and prioritize abstract human values (such as safety, health, or fairness) when making decisions.
7. **Principle of Double Effect (DDE)**:
   A moral principle that justifies actions that have both positive and negative outcomes, as long as the negative effects were not intended and the positive effect outweighs the harm.
8. **Theory of Mind**:
   A cognitive ability in humans (and potentially AI) to understand and predict the mental states and emotions of others, essential for ethical reasoning and empathy.
9. **Real-Time Ethical Decision-Making**:
   The ability of AI systems to process ethical dilemmas and make morally sound decisions within short timeframes, often in high-stakes situations (e.g., self-driving cars or medical emergencies).

Ch4:

ART Principles

1. **Accountability**

- **Definition:** Ability of AI systems to explain and justify their actions.
- **Short Note:** AI should be able to explain decisions and actions. If it makes a mistake, it should clarify why it happened. Design process and decisions should be documented.

2. **Responsibility**
   - **Definition:** Humans are ultimately accountable for AI systems' actions.
   - **Short Note:** AI is a tool made and used by humans. Developers and manufacturers are responsible for AI's behavior. Actions by AI should be traceable to design choices and user instructions.
3. **Transparency**
   - **Definition:** Openness about AI's design, data, and functioning.
   - **Short Note:** AI systems should be open about how they work, the data they use, and any biases they might have. This includes clear documentation and addressing biases.

Responsible AI and Its Societal Impact

1. **AI's Benefits and Risks**
   - **Definition:** Advantages and potential issues associated with AI.
   - **Short Note:** AI can improve accuracy and efficiency but also poses risks like privacy concerns and job displacement.
2. **Responsible AI Development**
   - **Definition:** Creating AI with consideration for ethical, legal, and societal impacts.
   - **Short Note:** Develop AI responsibly by including diverse perspectives and considering its broader impact on society.
3. **Importance of Education**
   - **Definition:** Educating people about AI to foster understanding and involvement.
   - **Short Note:** Educate the public and stakeholders about AI's impact to ensure informed participation in its development.

Design for Values

1. **Identifying and Interpreting Values**
   - **Definition:** Recognizing and understanding societal values relevant to AI.
   - **Short Note:** Identify key values and understand how they apply to AI design and functionality.
2. **Translating Values into Functionality**
   - **Definition:** Converting abstract values into specific AI features and rules.
   - **Short Note:** Turn values like fairness into concrete rules and actions for AI systems.

3. **Formalization and Traceability**
    - **Definition:** Documenting how values are linked to AI design choices.
    - **Short Note:** Ensure values, norms, and functionalities are clearly documented and traceable.

Responsible Development Life Cycle

1. **Aligning with Human Values**
    - **Definition:** Ensuring AI goals match societal values and involve stakeholders.
    - **Short Note:** Align AI systems with human values and involve people in setting AI goals.
2. **Explicit Interpretation of Values**
    - **Definition:** Clearly defining how values are understood and applied in AI.
    - **Short Note:** Specify how values like privacy are interpreted and implemented in AI systems.
3. **Ethical Reasoning Methods**
    - **Definition:** Outlining how AI handles ethical dilemmas.
    - **Short Note:** Define how AI should prioritize and handle conflicting values.
4. **Governance Mechanisms**
    - **Definition:** Structures for overseeing AI development and addressing issues.
    - **Short Note:** Implement governance to monitor AI decisions and handle potential problems.
5. **Openness and Data Provenance**
    - **Definition:** Transparency about AI's data sources and design choices.
    - **Short Note:** Document and share information about AI's data and design processes.

Ch5:

1. Ethical Actions & Challenges

- **Dennett's Requirements:**
    - Choose actions
    - Society agrees on benefits
    - Recognize and pick ethical actions
- **Challenges:**
    - Hard to define all actions & ethics
    - No universal agreement
    - Ethics evolve

2. Ethical Reasoning Approaches

- **Top-Down:** Follow pre-defined rules.
  - **Example:** Self-driving car prioritizing human lives.
  - **Limitations:** Too rigid, oversimplifies ethics.
- **Bottom-Up:** Learn from human behavior.
  - **Example:** AI learns from observing humans.
  - **Limitations:** Can inherit biases, depends on data quality.
- **Hybrid:** Mix of rules and learning.
  - **Example:** AI with built-in rules and learning from experience.
  - **Advantages:** Flexible, less biased.

3. Designing Moral AI

- **Value Alignment:** Determine and respect values across cultures.
  - **Example:** Respect cultural differences.
- **Ethical Background:** Choose and justify ethical theories.
  - **Example:** Balance privacy vs. safety.
- **Implementation:** Decide AI's autonomy and human oversight.
  - **Example:** Robots with guidelines and human intervention.

4. AI’s Ethical Status

- **Autonomy:** AI acts independently but lacks moral awareness.
  - **Example:** Robots can’t “understand” morally.
- **Robot Rights:** Too early for rights for machines.
  - **Example:** Robots don’t have feelings.
- **Patiency:** Focus on how AI is treated ethically.
  - **Example:** Align AI behavior with human values.
- **Transparency:** Clear responsibility and decision explanations.
  - **Example:** Developers explain AI decisions.
- **Distributed AI:** Complex systems can obscure accountability.
  - **Example:** Hard to pinpoint responsibility in networked AI.

5. Conclusion

- **Focus Shift:** Prioritize transparency and human values over just performance.
- **Ethical Frameworks:** Develop better accountability frameworks.
- **Ongoing Dialogue:** Collaborate continuously with all stakeholders.