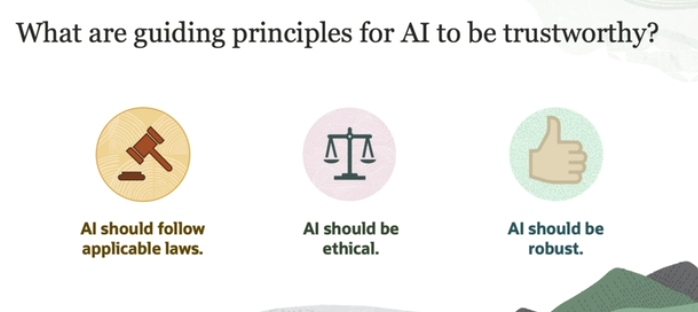


A person looking at a computer screen

Description automatically generated with medium confidence

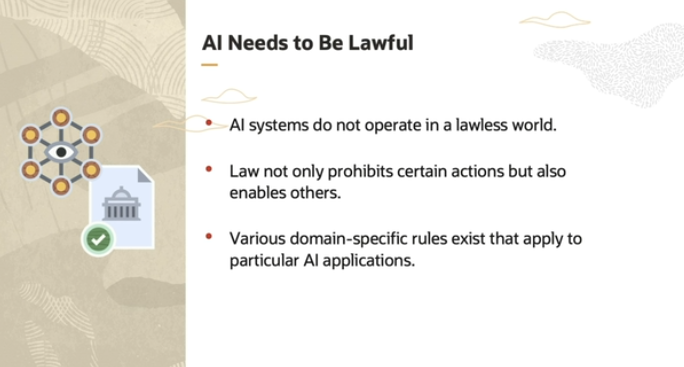
In this lesson, we will learn about responsible AI. We are increasingly using AI  in our day-to-day work. But can we trust AI? For example, do we trust self-driven cars completely, or do we accept AI driven disease diagnosis without a second opinion? For us to trust AI, it must be driven by ethics that guide us as well.



To ensure that AI is **trustworthy**, there are three main guiding principles:

1. **AI Should Be Lawful**: This means that AI systems must follow all the rules, laws, and regulations that apply. For example, AI must respect privacy laws or any legal restrictions related to how it can be used. Just like people or businesses need to obey the law, AI must do the same.
2. **AI Should Be Ethical**: AI should act in ways that align with human values and principles. This includes fairness, transparency, and respect for human rights. The idea is that AI should reflect what we, as humans, believe is right and wrong, avoiding bias or harmful actions.
3. **AI Should Be Robust**: AI needs to be strong and reliable both **technically** (working without errors) and **socially** (not causing harm to people or society). Even if the AI has good intentions or is designed for a positive purpose, it must be built carefully to avoid mistakes or unintended negative impacts.

In simple terms, trustworthy AI must follow the law, act ethically, and be safe to use in both technical and human aspects.

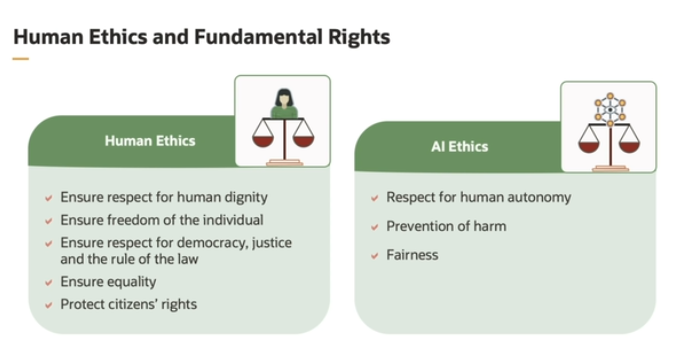


In simple terms, AI systems must follow rules and laws just like anything else in society. There are many **national and international laws** that apply to how AI is developed, used, and deployed today. These laws aren't just about stopping bad behavior; they also help protect important things, like **the rights of minorities** or **the environment**.

Additionally, there are two types of rules:

1. **General rules** that apply to all AI systems (like privacy or data protection laws).
2. **Specific rules** for certain areas, like healthcare. For example, **medical device regulations** apply to AI used in health-related technologies to ensure they are safe and effective.

So, AI operates in a world where laws exist to guide its development and use, helping to protect society and ensure fair treatment.



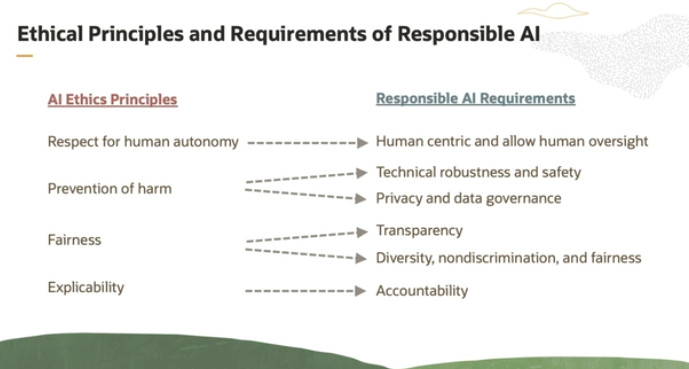
1. **Human Dignity**: This means that every person has a special worth that should never be taken away. AI systems should be designed to respect and protect people's **physical** (their body) and **mental** (their mind) well-being. In other words, AI should not harm or disrespect human beings.
2. **Freedom of the Individual**: People have the right to **freedom of expression** (to say what they think) and **privacy** (to keep their personal life and data safe). AI systems should support and protect these rights, helping to uphold **democratic processes** (like fair voting) and respecting individual choices. AI must not interfere with or damage democracy.
3. **Equality**: AI must treat people **fairly**. It shouldn't produce **biased** or unfair results that discriminate against certain groups of people. When we use AI, it’s important to make sure that everyone’s rights are protected and that no one is treated unfairly.

**How Do These Ethics Apply to AI?**

There are three key ethical principles for AI:

1. **Help Humans**: AI should be used in ways that benefit humans and should allow for **human oversight** (meaning people should be able to control or monitor it).
2. **Prevent Harm**: AI should never cause physical or social harm to individuals or society as a whole.
3. **Transparency and Fairness**: Decisions made by AI should be **clear, understandable, and fair**. People should be able to explain how and why the AI made a certain decision.

In short, AI ethics are about ensuring that AI **respects human dignity**, **protects freedom and privacy**, **promotes fairness**, and **avoids harm**, while making sure its decisions are always explainable and just.

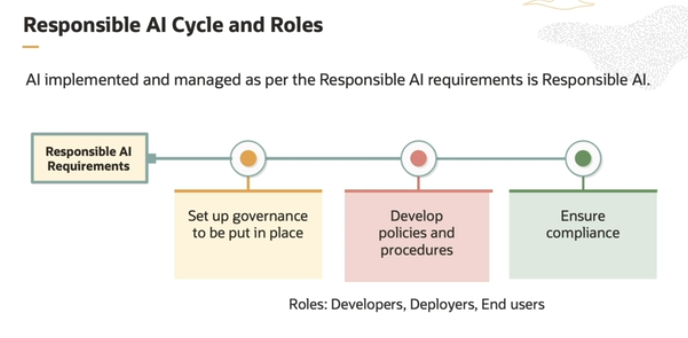


**Responsive AI** is AI that follows ethical principles, meaning it acts in ways that align with human values and fairness. Here’s how we can break it down:

1. **Human-Centric Design**: AI should be designed with **humans in mind**. It should always give people control and leave room for **human decision-making**. This ensures that people can oversee and control how the AI behaves.
2. **Safety and Security**: AI systems must be **safe to use** and protected from being misused. They should be built in a way that makes them **technically strong**, so they can’t be easily hacked or used for harmful purposes.
3. **Fairness**: AI systems must be **fair**, meaning that the benefits and costs of using AI should be shared equally. AI should not be **biased** or **discriminatory**. It should treat everyone fairly, without favoring certain people or groups over others.
4. **Explainability**: When AI makes decisions, it should be able to **explain** how and why it made those decisions. This is important for people who are affected by those decisions, both directly (like individuals) and indirectly (like society).

In short, for AI to be **responsible**:

* It must keep humans in control,
* Be safe and secure,
* Be fair and unbiased,
* And be able to explain its decisions.

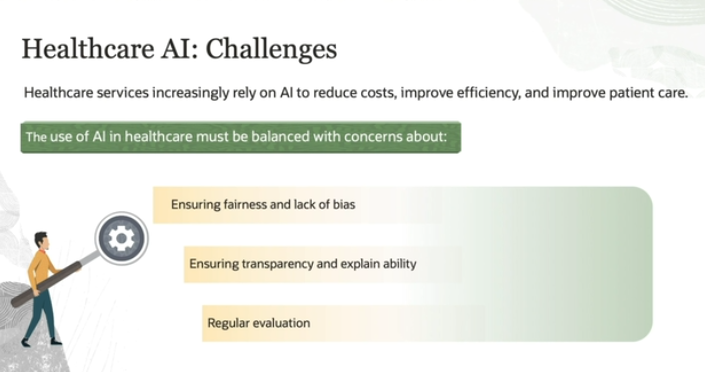


Let’s break down the cycle of **Responsive AI** in a simple way:

1. **Human-Centric Design and Human Oversight**:
   * AI should be built with **people in mind**, meaning that it should work in ways that benefit and protect humans. There should always be room for **human choice and control**, allowing people to monitor and manage the AI system. This ensures that AI is used responsibly and doesn’t act entirely on its own without human input.
2. **Safety and Security**:
   * AI systems need to be **safe** for people to use and must operate in a secure environment. They must be **technically robust**, meaning they are strong enough to handle challenges and are protected from **malicious use** (like hacking or being used for harmful purposes).
3. **Fairness**:
   * AI should treat everyone equally. The **benefits** of AI (like increased efficiency or new opportunities) and the **costs** (such as any risks or disadvantages) must be **shared fairly**. This means no one group should benefit more than others, and AI should not show **bias** or **discrimination** against certain people.
4. **Explainability**:
   * AI decisions should be **understandable** to the people affected by them. If AI makes a decision that impacts someone’s life, whether directly or indirectly, it should be able to **explain** how and why it made that decision. This helps build trust and accountability in AI systems.

**The Cycle of Responsible AI:**

* **Design**: Start by building AI with human values, making sure people remain in control.
* **Operate Safely**: Ensure the AI works in a safe and secure environment.
* **Fairness**: Make sure the AI benefits everyone equally and avoids bias.
* **Transparency**: The AI should explain its decisions so people understand them.



One of the key challenges in using AI in health is ensuring its fairness and lack of bias. AI systems are only as good as the data that they are trained on. If that data is predominantly from one gender or racial group, the AI systems might not perform as well on data from other groups. Another challenge in using the AI in health is ensuring that it's transparent and explainable.

AI systems often make decisions based on complex algorithms that are difficult for humans to understand. As a result, patients and health care providers can have difficulty trusting the decisions made by the AI. AI systems must be regularly evaluated to ensure that they are performing as intended and not causing harm to patients.