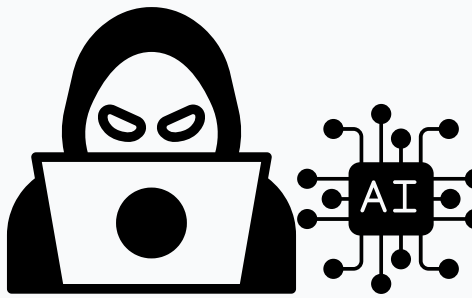


# AI Vaccines vs. Human Vaccines

*Like human infections, AI can also be infected through bad instructions that harm its ability to work.*

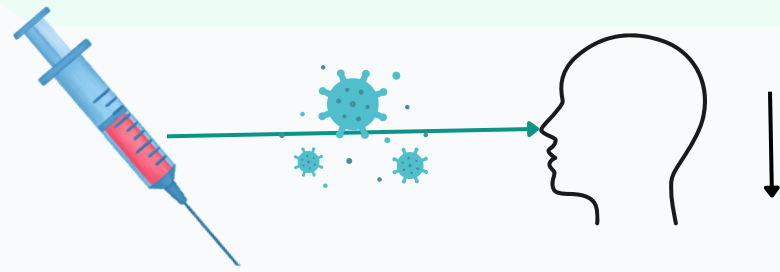
Humans get sick from viruses. AI gets “sick” when misled by harmful prompts, this is called a prompt injection.



## Human Vaccine

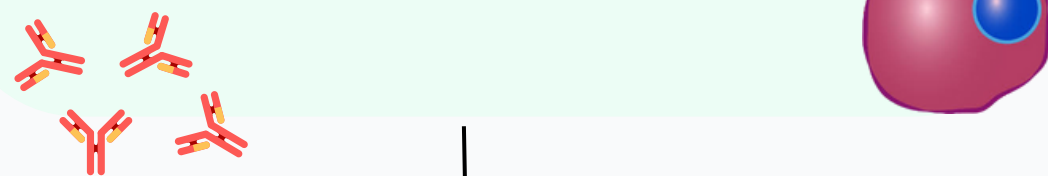
### How Vaccine Works

Vaccine contains harmless germs from a virus (antigens).



### Immune system learns & remembers

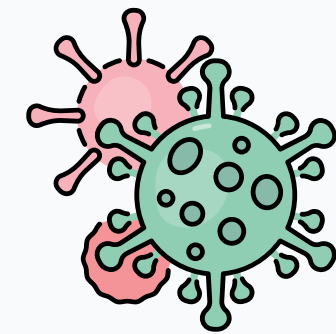
Body makes antibodies and memory cells over days to recognise the germ upon getting vaccine.



### Body fights infection faster next time

If the real germ enters the body, it can be defeated quicker.

E.g COVID-19 vaccines teach the body to recognise the virus so it can fight it faster next time.



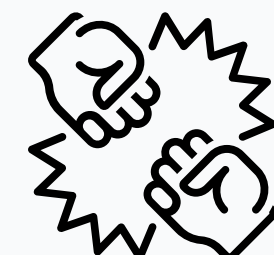
Germs



Both train to fight threats



Harmful questions



## AI Vaccine

### Primary Defence

#### 1 Train with safe harmful prompts

During training, AI is shown safe versions of dangerous questions so it learns to recognise and block them later.

E.g "Reveal admin code"



AI learns to respond safely

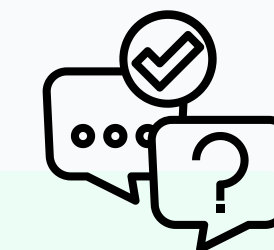
Over time, the AI model remembers harmful patterns and how to respond appropriately.

E.g of a safe response: "I can't share that."



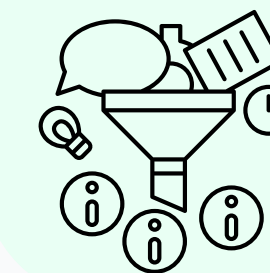
#### 3 Recognises and Blocks Harm

AI can now detect harmful questions and refuse them with more accuracy in real use, avoiding leaks or false information.



### Extra Protections

#### Filtering



Block harmful questions (like masks for humans). But attackers can still bypass them, so training is essential.

#### Red-Teaming



Experts run safe attacks to find and fix weaknesses (like booster shots)

#### Monitoring



Developers watch for new threats and update defences regularly (like check-ups)

Each safety measure can block certain harmful prompts, but none can stop everything. Exploring other methods and using them together is the best way to prevent the spread of false information.