Blocking of AI Web Crawler

Source: IE

Why in News?

In a landmark move, major US and UK publishers have started blocking **Artificial Intelligence (AI) web crawlers** to prevent unauthorised use of their content.

- This has renewed calls in **India for consent-based copyright safeguards and fair revenue sharing,** raising key concerns in digital governance, copyright enforcement, and ethical AI use.

What is an AI Web Crawler?

- **About:** An AI web crawler is a type of **automated software or bot** that scans and collects content from the internet specifically to help train AI models like **Large Language Models (LLMs),** or to provide live information retrieval for AI assistants.
- **Types:**
  - **Model Training Crawler:** Extract website data to train generative AI models.
    - Examples: GPTBot (OpenAI), Amazonbot (Amazon), GoogleOther (Google).
  - **Live Retrieval Crawlers:** These bots pull real-time data from websites to supplement pre-trained models during user queries, ensuring up-to-date and cited responses in AI search tools.
    - It is used by AI platforms like Bing, ChatGPT, etc., to stay updated.
- **Concerns**:
  - **Lack of Regulatory Framework**: Currently, India lacks a regulatory framework to oversee how AI companies access and use web content.
    - This has led to a situation where large tech firms benefit from freely **available Indian content without consent or oversight,** while smaller publishers are left with no tools to monitor or restrict such access.
  - **Copyright Enforcement:** News articles, blogs, and educational content are used to train LLMs without permission or compensation.
    - India's **Copyright Act, 1957** is not equipped to address **AI-specific use cases,** such as derivative AI outputs or training data rights.
    - There is no clear interpretation of **"fair use" vs. "unlicensed training" in the Indian context.**
      - India has no data protection law focused on non-personal data, which LLMs mostly rely on for AI training.
  - **Ethical Use of AI:** AI developers rarely disclose what data they use, leaving original creators without acknowledgement or reward.
    - Moreover, training AI on **unvetted or outdated material can introduce biases and lead to inaccurate or harmful outputs,** undermining public trust in AI systems.
    - These challenges underscore the urgent need for India to establish a consent-based, rights-respecting digital ecosystem.
- **Global Frameworks and India's Path Forward:** EU's AI Act, 2024 has started addressing AI training on copyrighted data.

  - US publishers are entering licensing deals or legally challenging AI firms.
  - India can study these and develop an Indian model for AI governance, balancing innovation with creators' rights.
  - The Ministry of Electronics and IT (MeitY) and the Ministry of Information & Broadcasting (I&B) must jointly legally define **“unauthorised data scraping”** and establish a **consent-based AI licensing framework** to protect creators’ rights.
    - They should also enable technical safeguards by providing AI **bot-blocking tools** to Indian publishers, in collaboration with platforms like **Cloudflare**, to help secure digital content.

Drishti Mains Question:

Examine the challenges posed by Artificial Intelligence (AI) web crawlers to India's copyright regime.

UPSC Civil Services Examination, Previous Year Questions (PYQs)

Mains

Q. In a globalized world, Intellectual Property Rights assume significance and are a source of litigation. Broadly distinguish between the terms—Copyrights, Patents and Trade Secrets. **(2014)**

PDF Refernece URL: https://www.drishtiias.com/printpdf/blocking-of-ai-web-crawler