



# Brief: AI Training Data

---

Presented by Dr. Kira Allmann, JCOTS Chief Policy Analyst

# What is training data?

---

Training data is computer-readable information (datasets) used to teach artificial intelligence (AI) models to identify patterns, make predictions, or generate new content.

Some of the most commonly used training datasets are scraped from the internet.

# Policy Issues

---

- Copyright infringement
- Consent
- Discrimination
- Exposure to harmful content



*The AI “black box”*

# Proposed Policy Solutions

---

## **Transparency:**

- (1) Make training data scrutable to the public, researchers, and/or oversight and auditing bodies;
- (2) Provide the public with clear information about how their data could be used in AI training alongside options for opting in or out.

# Considerations & Critiques

---

## **The limits of disclosure**

- Understandability
- Actionability
- Verifiability

## **Meaningful options and redress for the public**

- Opt-out and disclosure go hand-in-hand
- Right of action and oversight

## **Protection of trade secrets**

- Balancing disclosure with commercial interests

Thank you