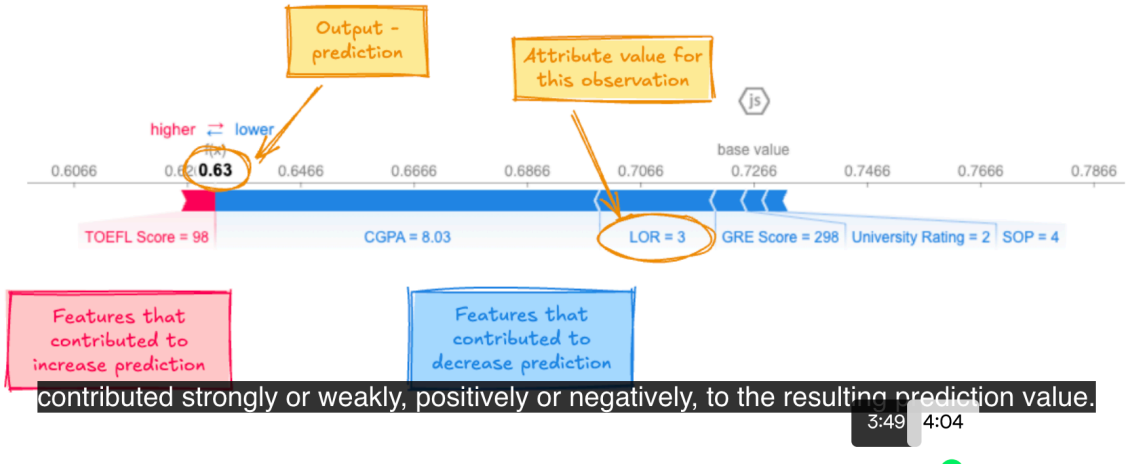
AI Transparency - XAI

SHAP	LIME	Decision Trees
Highly complex model with difficult to understand decision-making	Extremely precise predictive model, requiring specific explanation for individual predictions	Simple model with a need for clearer understanding of decisions.

XAI tools: feature importance

SHAP (SHapley Additive exPlanations)



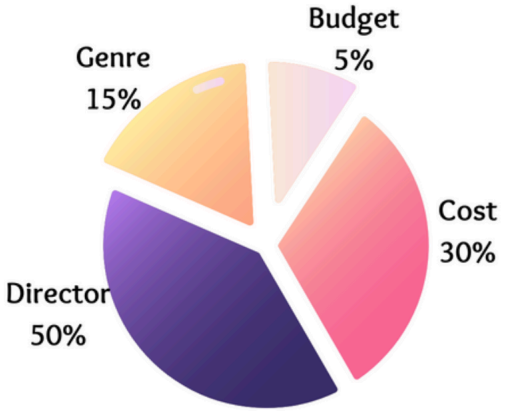
contributed strongly or weakly, positively or negatively, to the resulting prediction value.

3:49 4:04

Explainable AI 50 XP

SHapley Additive exPlanations (SHAP)

- SHAP: A detective of AI, revealing **feature importance**
- SHAP in Action
  - Director: 50%
  - Cast: 30%
  - Genre: 15%
  - Budget: 5%



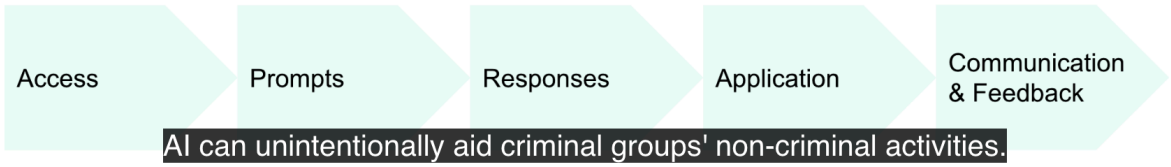
prediction, the cast was 30% responsible, the genre was 15% responsible, and the budget was 5% responsible.

Detection and prevention

Key usage principles

- Human-in-the-loop
- Harm prevention
- Continuous monitoring

Points of Detection and Prevention



AI can unintentionally aid criminal groups' non-criminal activities.

Use the right tool for the job

Discriminative Models

- Predict tomorrow's weather
- Categorize books
- Determine if a picture is a puppy or a bagel

Generative AI

- Write code for a website
- Answer unique customer service questions
- Draw a picture of a cat scuba diving

Artificial General Intelligence

- Complete traditionally human jobs