AI and ML Ethics

By: Winston Essibu and Einstein Essibu

Topic for discussion: Ethics about machine learning models and Ai models , in particular the issues with transparency and explainability as well as the issues with manipulation and misinformation.

Introduction to AI Ethics

AI: the ability of machines to perform tasks that typically require human intelligence, such as learning, problem-solving, and decision-making.

- AI and ML ethics
- What is AI and ML ethics: the moral principles and practices that guide the development, deployment, and use of artificial intelligence and ML systems to ensure they are designed and used in a responsible and beneficial manner

Transparency and Explainability

- What is Transparency in AI and ML ? open, understandable, and accessible
- Why is it important ?
  - Enables accountability and oversight
  - Helps detect and mitigate bias or unfairness
  - Builds trust with users and stakeholders
  - Is often required by regulations (like the EU AI Act)

Cont. Transparency and Explainability

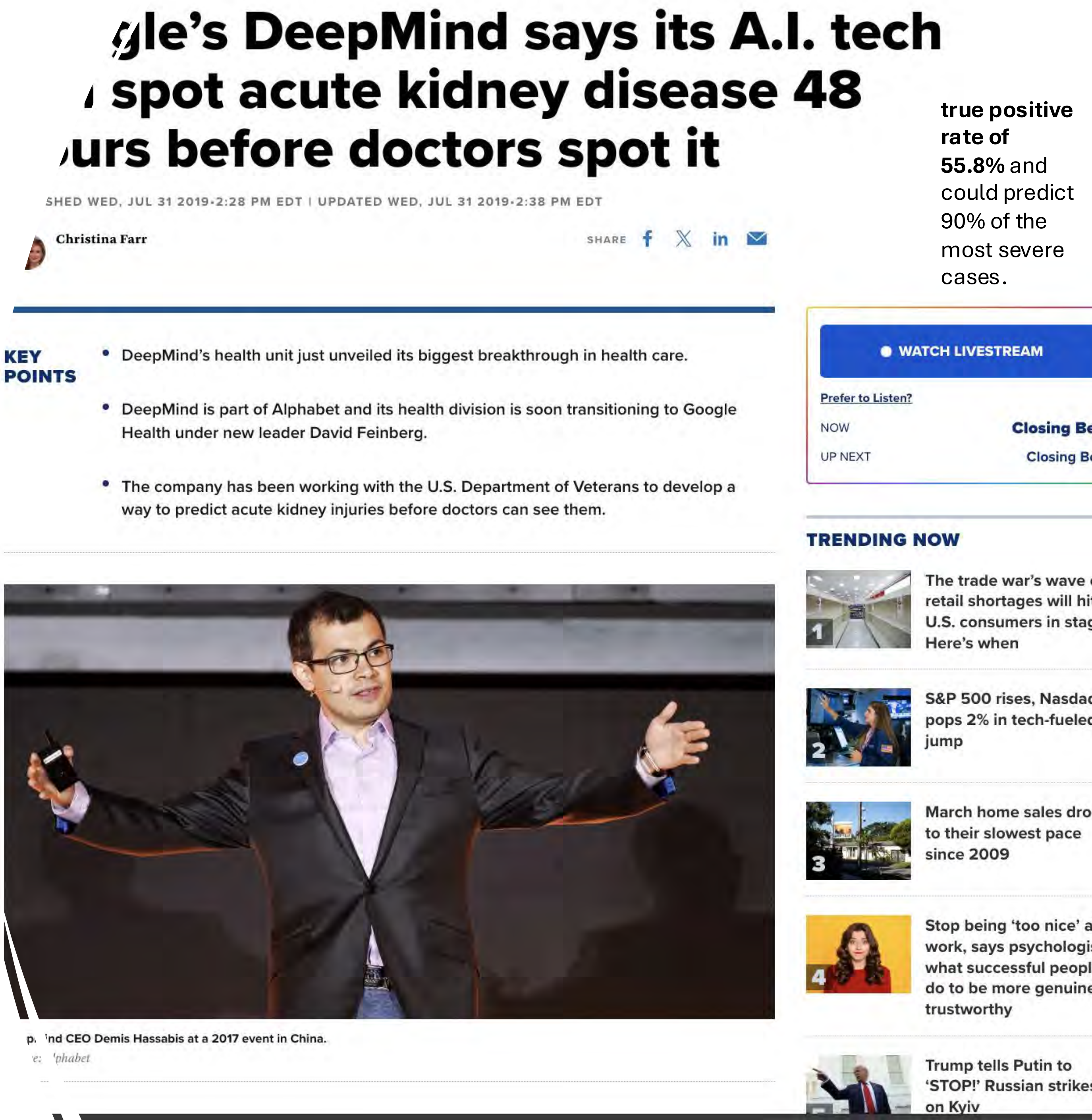
What is Explainability? describe, in human terms, how an AI system reached a certain outcome.

Why Explainability

- User Understanding
- Trust and Fairness
- Error Detection

Case Study

- The AI was a **black box**
- Doctors found it hard to **trust or act on the predictions**
- In healthcare, every decision needs to be justifiable



gle's DeepMind says its A.I. tech spot acute kidney disease 48 urs before doctors spot it

SHED WED, JUL 31 2019·2:28 PM EDT | UPDATED WED, JUL 31 2019·2:38 PM EDT

Christina Farr

KEY POINTS

- DeepMind's health unit just unveiled its biggest breakthrough in health care.
- DeepMind is part of Alphabet and its health division is soon transitioning to Google Health under new leader David Feinberg.
- The company has been working with the U.S. Department of Veterans to develop a way to predict acute kidney injuries before doctors can see them.

p ind CEO Demis Hassabis at a 2017 event in China.

e: phabet

TRENDING NOW

1 The trade war's wave retail shortages will hi U.S. consumers in stag Here's when

2 S&P 500 rises, Nasda pops 2% in tech-fuele jump

3 March home sales dro to their slowest pace since 2009

4 Stop being 'too nice' a work, says psychologi what successful peopl do to be more genuine trustworthy

Trump tells Putin to 'STOP!' Russian strike on Kyiv

true positive rate of 55.8% and could predict 90% of the most severe cases.

Solutions

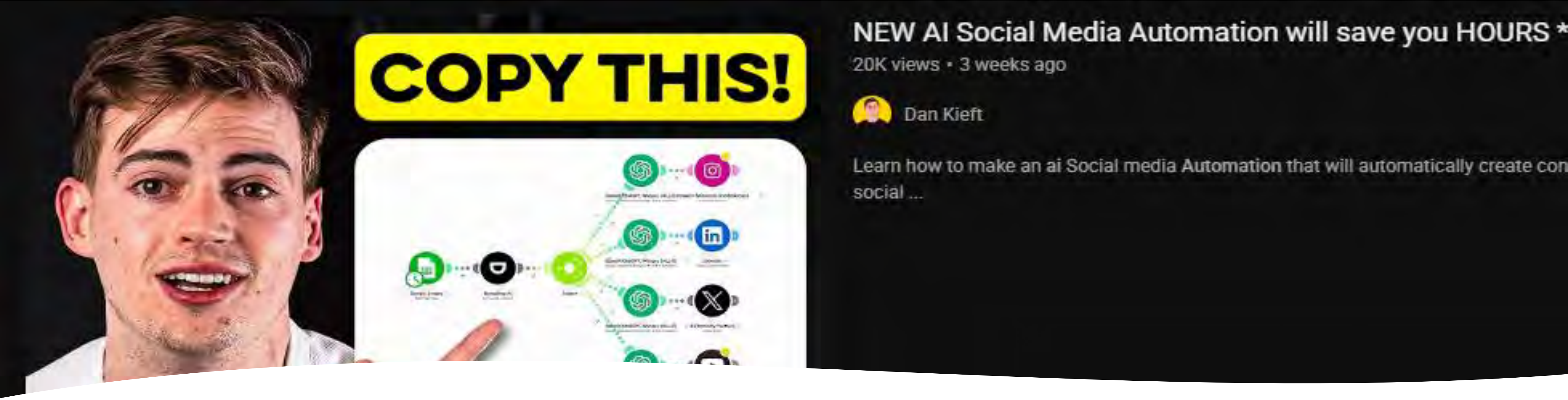
2. Post-Hoc Explanation Tools (Explain After Training)

- a. LIME (Local Interpretable Model-Agnostic Explanations)
- SHAP (SHapley Additive exPlanations)

Use Interpretable Models When Possible

- Decision Trees, Logistic Regression, Rule based System

The Evolving Landscape of AI-Enabled Misinformation and Manipulation



NEW AI Social Media Automation will save you HOURS *

20K views • 3 weeks ago

Dan Kieft

Learn how to make an ai Social media Automation that will automatically create con
social ...

AI Misinformation in Social Media Automation

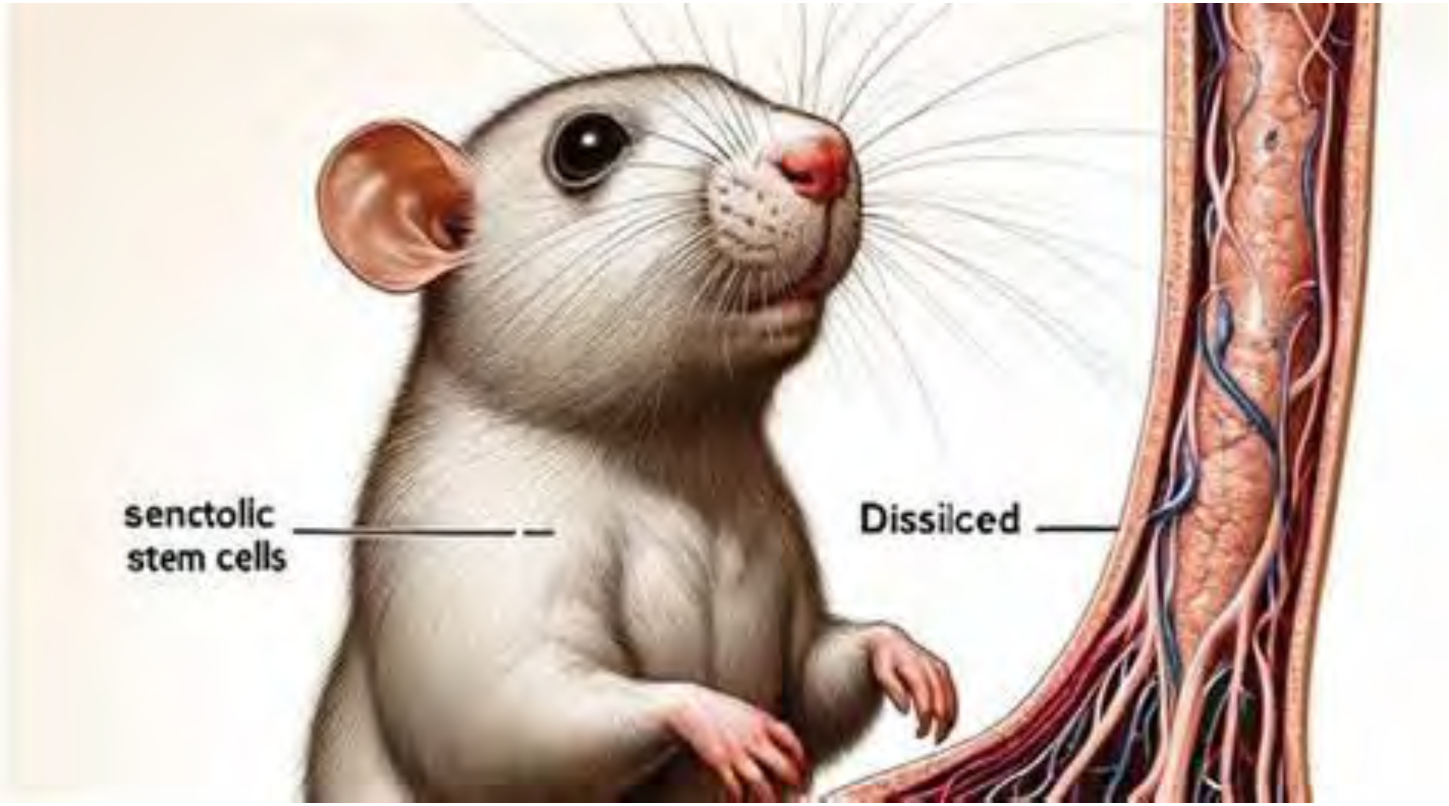
- Platforms reward consistency
  - 3+ video or posts s a day to exploit engagement metrics
  - @voidstomper achieve 20M+ views with minimal editing
- Dedicated Tools created for converting text prompt to Youtube videos
  - Pictory.ai, Vizard.ai, Fliki
  - N8N Automation
  - YouTube Tutorials
- Rise of AI-generated content farms
  - Create as much videos as they can
  - Automate the distribution of farms

AI Manipulation in Scamming

- **Hyper-Realistic Impersonation via Deepfakes**
  - Voice/Video Cloning
  - fraudsters mimicked a CFO's voice to trick an employee into wiring $25M
- **AI-Optimized Phishing**
  - Personalized Scam Content
  - Scanned victims' LinkedIn job histories to fake "HR onboarding" requests
- **Algorithmic Evasion**
  - AI adapts scam patterns in real-time to bypass fraud filters
  - LLMs are instructed alter phishing email syntax hourly to avoid keyword detection

AI Misinformation and Manipulation in Scientific Research

- An Industry that rewards Publication Frequency
- Fabricated citations ("hallucinations")
- AI Generated Pictures
- Plot or scatter-plots that never existed



Solutions and recommendation

- AI detection algorithms outperforming human judgment (74% accuracy)
- Add "fact-check" to search queries
- Be suspicious of generic website titles
- Verify information across multiple sources

Thank you!!

Any Questions ?