How can we ensure that evals are not gamed?

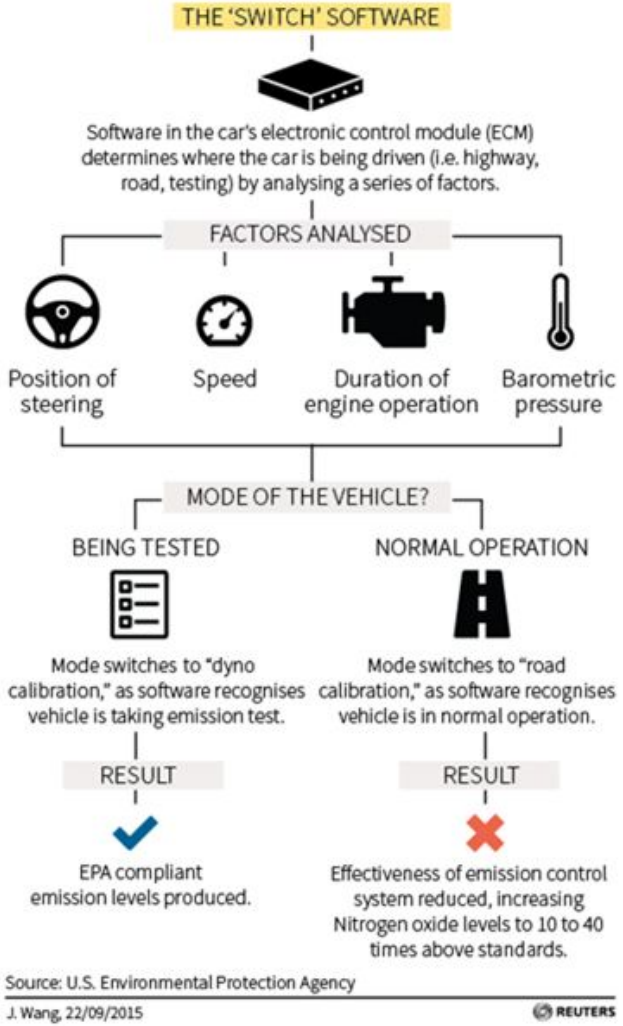
An analysis drawing on the Volkswagen gas emissions test case study

AI Governance Research Sprint
London EA Hub, 9 March 2024

Rebecca Hawkins, Vinay Hiremath, David Varga

Volkswagen Emissions Scandal

- In 2015, the US Environmental Protection Agency (EPA) found that in over 590,000 diesel motor vehicles, Volkswagen had violated the Clean Air Act
- The vehicles were equipped with “defeat devices” in the form of a computer software
- This was designed to cheat on federal emissions tests.
- A defeat device is one that bypasses or renders inoperative a vehicle’s emission control system
- Essentially, software of this kind is designed to detect when the vehicle is undergoing an emissions test
- It then turns on full emissions controls during the testing period
- In the course of normal driving, the effectiveness of such devices is reduced



VW leadership and decisions

- Winterkorn, named CEO of VW in 2007, announced a bold plan to make VW the largest and greenest car company by 2018
- Diesels traditionally failed to meet US diesel emissions limits (stricter than in the EU), and VW claimed to solve this with its “clean diesels” increasing US diesel sales 150% in 4 years
- Bypassing testing with a defeat device was mentioned in a meeting with engineers and executives, with one member warning against the potential reputational harm if exposed but this was ignored

Discovery of emissions defeat code

- The International Council on Clean Transportation (ICCT) commissioned university researchers at the Center for Alternative Fuels, Engines, and Emissions (CAFEE) to understand VW clean diesel tech, which had claimed to solve the NOx and other emissions problems
- Known existing methods were more cumbersome, requiring large devices and/or tanks that needed frequent refilling
- Researchers used mobile measurement to test real-world use, found anomalies that were initially thought incorrect, eventually concluded different real-world behavior
- California Air Resources Board (CARB) was particularly focused on air quality and smog (often caused by NOx) and after hearing of the CAFEE work commissioned a year-long study

Effects of emissions cheating

- Following US investigations, VW officials estimated liability at $20 billion and decided to cover it up, publishing a sham software fix, destroying thousands of documents, and ditching mobile phones
- Eventually admitted fraud after being threatened with non-certification for all 2016 vehicles
- Many manufacturers tuned engines for testing, but VW explicitly enabled different behavior when testing
- Emitted up to 40x the US limit for NOx, which resulted in about 45,000 DALYs lost
- In response, the EPA and other US agencies increased future testing, extended the certification process, issued fines, mandated buybacks and compensation
- European regulators, facing widespread noncompliance and much higher diesel passenger car numbers than other regions, weakened testing for several years afterwards (allowing NOx 110% above the legal limit)

An example for process-based testing

- EPA could only test ~15% of vehicles due to limited resources, and this concern remains with AI evals, with less resources in evals than training
- PCI compliance is one analog to improve internal processes for software
  - Credit/debit card fraud was common since the late 1990s, and issuing banks who paid the cost of this fraud started by publishing independent security standards for those handling payment data, eventually became Payment Card Industry Data Security Standard (PCI DSS)
  - Higher level of certification requires an audit by an independent external assessor
  - Requires industry best practices, annual security training, providing test cases to external auditor, maintaining inventory of software/documentation
  - Also mandates separation between development/production staff, and segregation of functions **indispensable to org and subject to abuse** to minimise mistakes/cheating

Substantially Different Outcomes Depending on Jurisdictions

	USA	**Germany + EU**
Vehicles	500k	8.5m
Outcomes	● 20 billion USD buyback program offered to owners (covers buyback or modification) ● 2.8 billion USD criminal penalty ● 4.7 billion USD for clean air projects and consumer education programs	● Fines up to 25,000 EUR for non-compliance (varied by car model and emissions excess) ● Mandatory software updates for millions of vehicles (implementation timeframe varied)
Systems	● Robust legal process forcing VW to cooperate (class actions combining individual claimants, etc.)	● No criminal prosecutions in Germany ● There's no criminal liability for corporations ● There's no statute barring a criminal conspiracy, no relevant criminal clean air law, and no law against lying to regulators or investigators

Possible Analogies with AI Model Evaluations

Volkswagen Case Study	**AI Model Evaluations**
Feasibility of relatively low-cost independent testing (in labs and real-world) by universities and governments	AI models can be evaluated with access to the model
Portable emissions testing equipment avoided VW measures that applied to in-lab (treadmill) testing	Running evaluations in a way that the model doesn't recognise it's being tested
"Defeat device" (code within the vehicle "ECU" (Electronic Control Unit))	E.g. rewarding the model for evasiveness during the training runs to be applied during evaluation
Incentive against enforcement: Diesel cars had much greater integration in the economy in EU vs USA	Important to track integration of advanced AI in critical sectors of the economy
Management Scheming	Same
Unrestricted Deception Process Control	Uncertain. In theory could be same, but possibly more resistance to regulator access to commercial info
Whistleblowing: none ex-ante, but helped during investigation	Further investigation likely to show lessons for AI

Further avenues for research

- Research large established regulators (SEC, EPA) and their regulatory net-impact (benefits, and downsides)
  - Building analogs to more effective enforcement agencies in the VW diesel scandal
- Evaluation and auditing of processes rather than final product evaluations
  - Existing frameworks for process auditing (PCI, etc.)