



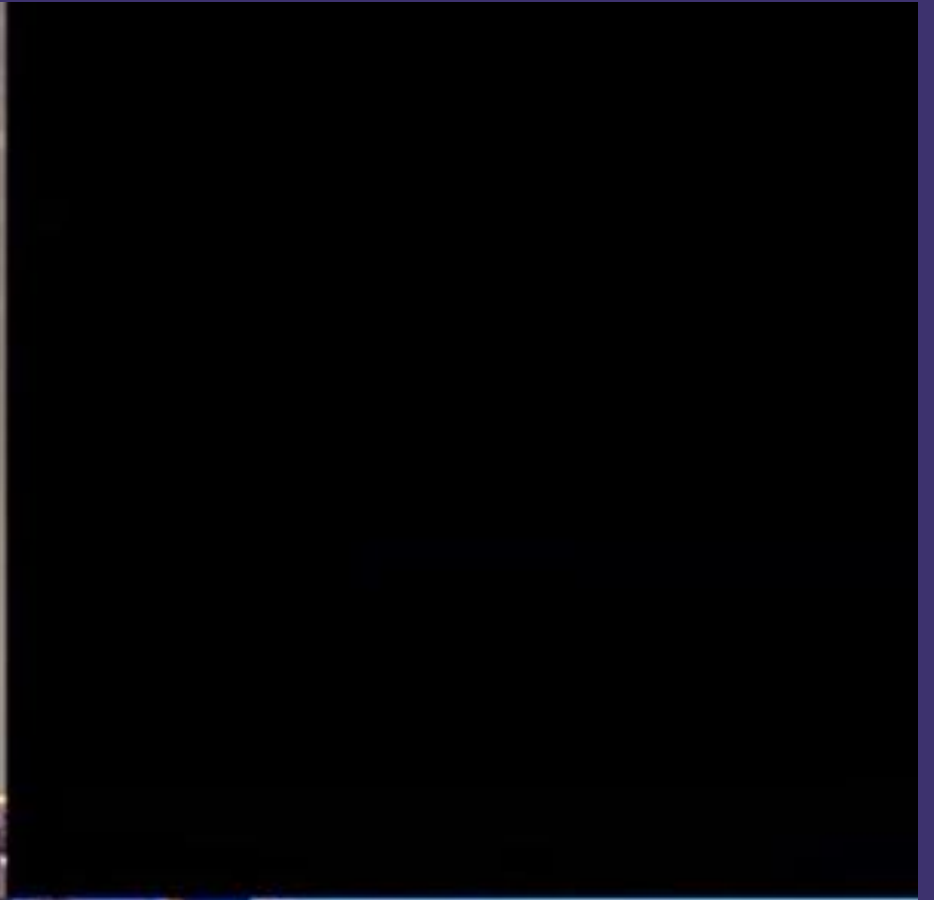
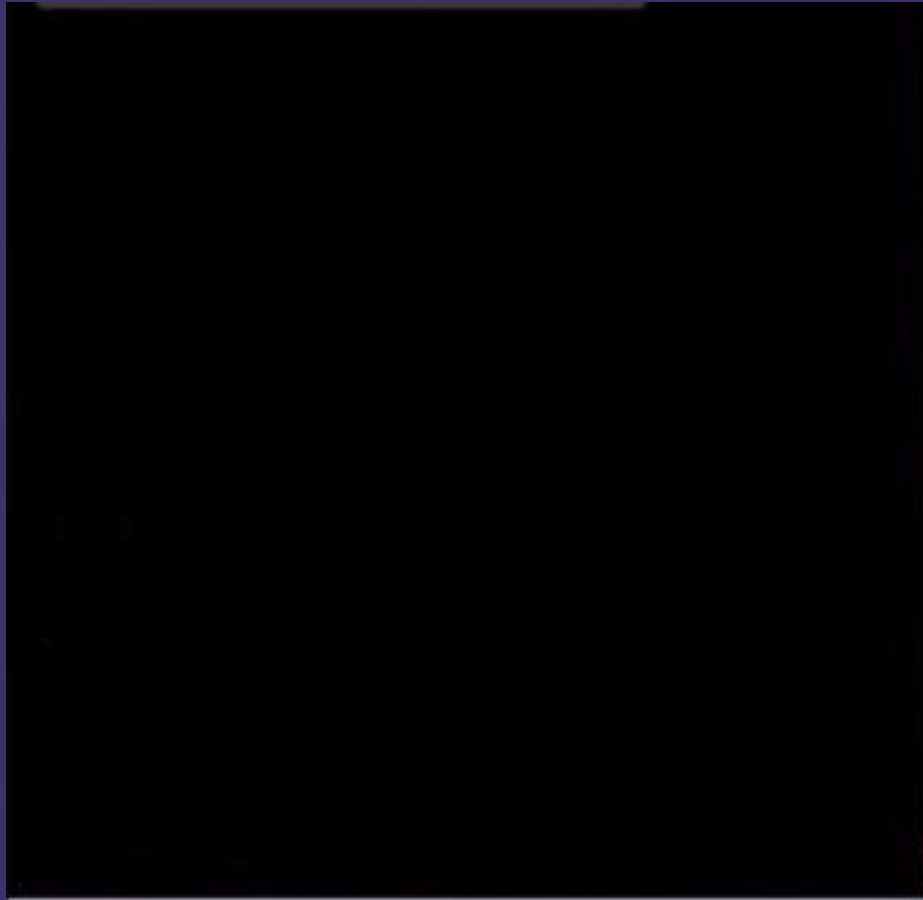
AI IN COURT

DR. ANDREW CHEN



Is evidence derived from AI systems admissible in court?

- It depends... but on what?
- Courts have produced guidance on generative AI
(mostly focusing on AI-written submissions and citations)
- AI-derived evidence may come from genAI text, but could also come from automated analysis or enhancement of images and audio, pattern recognition, scene reconstruction, matching biometrics, inferring behavioural analytics, etc.
- Questions of validity, reliability, bias, explainability, fairness
- Using AI to falsify evidence is a consideration, but a separate issue
- A series of mock trials to test these arguments practically
 - Hosted by the High Courts in Christchurch, Auckland, Wellington
 - Real judges, senior counsel, student counsel, expert witness
 - Fake facts and technology (but inspired by real systems)
 - Court provided with a Technical Expert Witness statement about the tool



MOCK TRIAL

Disclaimer: All characters are fictional;
any resemblance purely coincidental.

**The Law
Association
OF NEW ZEALAND**

What is the nature of evidence?

- Presumption of innocence
- Classification by:
 - Form (real/physical, documentary, testimonial)
 - Content (direct, circumstantial, corroborative)
- Evidence must have relevance, materiality, and competence
- Admissibility vs Weight
- Different standard of proof between criminal and civil proceedings
- Different expectations of discoverability and inspection between sides
- Evidence must provide facts, not opinions
 - (except for expert witnesses within their field)
- Overarching need to maintain “an effective and credible system of justice”
 - The nature of the evidence and the seriousness of the offence are relevant factors
 - “If it’s dirty try thirty” (s30 Evidence Act 2006)

R v Popper

- MDMA pills with a distinctive marking (“Goodnight Kiwi” logo)
- Sold by a single supplier via the dark web
- Police use the (hypothetical) “TraceLex” tool to match text from the dark web posts against an OSINT database
 - Using “lexical analysis” to match writing styles
- Matches found two Reddit accounts
 - One included a partial reflection of the photographer’s face
 - One included a screenshot of a utilities invoice
- Upon review by intelligence analyst, highly likely that Mr Popper is behind Reddit accounts and dark web posts
- Applied for Surveillance Device Warrant, made a controlled purchase, photographed the offender, applied for Search Warrant, seized drugs and a pill press with the logo at his house
- Clearly the guy did it... but is all of the evidence fruit of the poisoned tree?
- Should the Surveillance Device Warrant have been issued based on the TraceLex evidence?



Mr Phillip Popper (AI-generated)

DRUG INTERDICTION BRANCH

Subject Profile

Phillip POPPER

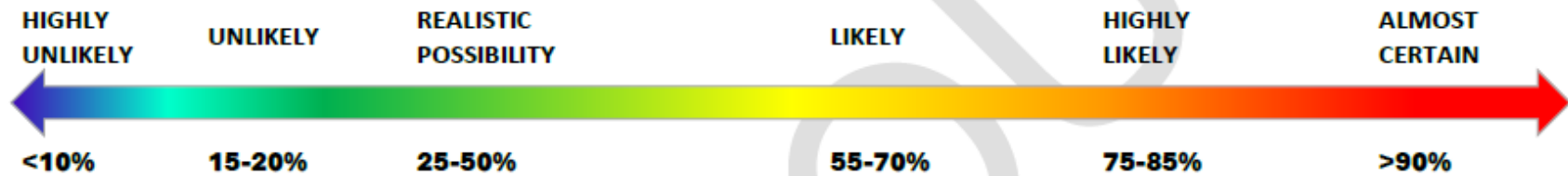
9 August 2024



Probabilistic Language

Probability Statement	Qualitative Statement	Percentage Probability
ALMOST CERTAIN	A scenario or explanation that has only a remote chance of not occurring or not being currently accurate.	>90%
HIGHLY LIKELY	A scenario or explanation that has only a small chance of not occurring or not being currently accurate.	75 - 85%
LIKELY	A scenario or explanation that is more likely than not to occur or to be currently accurate.	55 - 70%
REALISTIC POSSIBILITY	A scenario or explanation that has a realistic chance of occurring or being currently accurate, but which does not outweigh all other alternative possibilities and explanations.	25 - 50%
UNLIKELY	A scenario or explanation that is plausible, but which has only a small chance of occurring or being currently accurate.	15-20%
HIGHLY UNLIKELY	A scenario or explanation that only has a remote chance of occurring or being currently accurate.	<10%

Probability Yardstick



Designing TraceLex

- Text-based AI system that matches writing samples to determine common authorship
 - Handwriting analysis, but for word/text patterns, automated at scale
 - GenAI Detectors that look for word/text patterns that indicate an AI author
- Specific methods are “commercially sensitive” and expert witness is guessing
 - *Loomis v Wisconsin* [US WI 2016] – “trade secret” COMPAS Algorithm
- Probability-based technologies that inherently produce variable results
 - *Matter of Weber* [US NY 2024] – Microsoft Copilot accounting calculations not reproducible
 - Cloud-based tools could be updated at any time and produce different results – lack of repeatability
- Non-linear confidence scores and thresholds (that can be misinterpreted as percentages)
 - *Bridges* [2020 EWCA Civ 1058] – Live FRT use by South Wales Police had unclear threshold policy
- Open Source Intelligence (OSINT) and collection of information from “public online spaces”
 - GDPR Article 9 only allows processing of sensitive data if “manifestly made public”

Technical Expert Witness

- Asked to evaluate and explain how the TraceLex software works
- The company does not publicly disclose its methods
 - No primary information source to rely upon
- Secondary information includes:
 - News sources speculating how the system works
 - Direct experiments conducted by the expert witness
- Speculation that in addition to lexical analysis, timeprinting, profiling, and image recognition used as well
- Three (hypothetical) experiments:
 - Writing samples from academic researchers: 8/8 correctly identified, confidence scores >80
 - Essays from Police recruits: 2/12 correctly identified, confidence scores <30
 - Rugby league group chat: 7/10 correctly identified, confidence scores 60-80
 - Shows that the amount of publicly available writing makes a difference (OSINT)
 - Shows that people write in different tones dependent on context (e.g. essays vs social media)

Legal Analogies

- **Anonymous or confidential informants**
- Cannot be meaningfully tested and difficult to assess the reliability at the time the info is provided
- Need to establish that the “assertions are solidly grounded in more than mere suspicion, rumour, or gossip”
- **Scientific equipment**
- Most AI outputs *could* in principle be produced by humans, but AI does it at scale and automatically
- Is it any different to the use of any other machine? Database search tools, speed cameras, breathalysers
- The Courts accept the outputs of these devices when the inner workings can be explained by experts
 - Testing, peer review, documented error rates (and no material bias), and “general acceptance”
- AI generally lacks explainability due to the black-box nature of these tools
- **Hearsay**
- AI outputs cannot be directly tested for their validity during cross-examination
- Hearsay only admissible if there is reasonable assurance that the statement is reliable
- Some case law that hearsay cannot apply because there is no human declarant

S30 Evidence Act

- A balancing test to allow for improperly obtained evidence to still be held admissible
- Was the use of TraceLex reckless or done in bad faith?
- Were there any other investigatory techniques available that would not have breached the defendant's rights?
- How was the impropriety balanced against the seriousness of the offence?
- Would exclusion of the evidence compromise an “effective and credible system of justice”?
 - “excluding highly probative evidence of serious commercial MDMA dealing would undermine public confidence, because it would let a major supplier escape justice”
 - “credibility would be compromised if the courts were seen to be rubber stamping invasive AI-driven surveillance tools by allowing Police to keep the evidence”
- s30 is not a panacea – once it has been used, it can't be used again in good faith
 - The evidence is still “improperly obtained” so the process/method cannot be relied upon again

Judgement

- TraceLex not solely relied upon, human-in-the-loop checks and verification occurred
- Arguable if these checks are meaningful if the human operators are not experts
- Two judges ruled that the evidence was collected unlawfully
but it would be admissible under s30 of the Evidence Act
- One judge ruled that the evidence was collected lawfully and therefore admissible
- Audiences were split 50/50 on admissibility in two trials, but 90% voted inadmissible in one trial (with some judicial influence on the audience vote)
- Use of the tool by Police was not done in bad faith – but they likely couldn't do it this way again
- Difficulty in establishing accuracy and reliability thresholds
- The information was ultimately accurate in this case, but it may not be in another case
- An insight into the battlegrounds we might face through the Courts
- **The Courts must balance innovation with fundamental rights**

Event Recording (AKL)



<https://www.youtube.com/live/0ofLQSC85hU>

Document Pack



<https://tinyurl.com/ai-in-court-pack-2025>