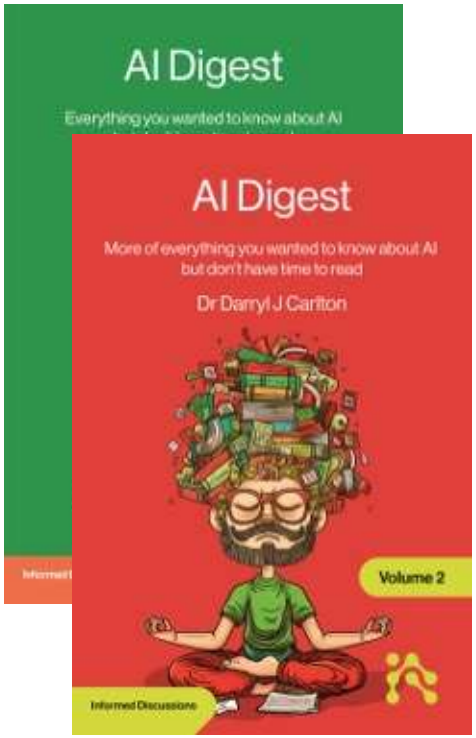
Responsible AI in a Turbulent World

Embedding Governance in Systems, Not Slogans

Speaker



Dr Darryl Carlton

and Simba

G.Dip C&IS., MBIT, M.Comm., PhD (IS)., PhD (Law), GAICD., MACS (Snr) CP

Today's Journey: From Failure to Framework

ACT I: THE CRISIS

Real Lives, Real Failures

- Jenny Miller & 433,000 victims
- Seema Misra & 736 wrongful convictions
- The implementation gap killing innovation

ACT II: THE DIAGNOSIS

Why Governance Fails

- Three digital empires, divergent visions
- Checkbox compliance vs real governance
- 95% failure rate despite frameworks

ACT III: THE FRAMEWORK

8 OECD Principles Decoded

- From theory to practice
- Real cases, real lessons
- What actually went wrong

ACT IV: THE SOLUTION

5 Practices That Work

- Test with dirty data
- Shadow mode deployment
- Clear stop-work authority

Your Choice: Theatrical Compliance or Genuine Governance

Garbage In - Gospel Out

"In 2019, Jenny Miller, a 65-year-old grandmother from Brisbane, received a letter that would destroy her life."

The algorithm had averaged her casual work incorrectly, claiming she owed $4,000 she never took.

She was one of 433,000 Australians wrongly pursued for $1.73 billion.

Source: Royal Commission into the Robodebt Scheme Report (2023)

The Three Digital Empires

Each imposing its vision of how AI should govern our lives

EUROPEAN UNION

'Protect the Citizen'

- Rules first, innovation follows
- AI Act: 113 articles
- **Penalties up to 7% of global revenue**

UNITED STATES

'Protect the Innovator'

- Innovation first, rules follow
- 'Move fast and break things'
- **1,080 state AI bills in 2025**

CHINA

'Protect the State'

- AI for nation building
- Social outcomes priority
- **State-directed innovation**

Source: Anu Bradford, Digital Empires (2023)

The Real Villain: Checkbox Compliance

95%

of AI pilots fail to deliver measurable impact

10 of 160

AI ethics guidelines include enforcement

'It is shocking how many members of responsible AI teams are being let go at a time when, arguably, you need more of those teams than ever'

- Andrew Strait, Ada Lovelace Institute (Financial Times, March 2023)

The OECD AI Principles

What They Actually Mean in Practice

1 Human Wellbeing

2 Human Values

3 Transparency

4 Robustness

5 Accountability

6 Human Oversight

7 Privacy

8 Sustainability

Principle 1: Human Wellbeing

Case Study: Amazon's AI Recruitment System

THE THEORY

'AI should benefit humanity'

WHAT HAPPENED

- Penalised CVs containing 'women's'
- Downgraded women's college graduates
- **Learned male = success predictor**

KEY LESSON: **Your training data embodies every historical bias, prejudice, and systemic inequality**

Principle 2: Human-Centred Values & Fairness

Case Study: Algorithmic Mortgage Discrimination

$250-500 million

Extracted annually in excess interest from minority communities

UC Berkeley Research Finding:

- Black and Latino borrowers paid 7.9 basis points more for purchases
- 3.6 basis points more for refinancing
- Same credit profiles, government-guaranteed loans
- **Algorithms learned to identify and exploit those who shop less**

Principle 3: Transparency & Explainability

Case Study: Dutch Childcare Benefits Scandal

40,000+ families affected by 'black box' algorithm

The Hague District Court Ruling:

'No transparency about the risk model used and the indicators from which this model consists' - Violated European Convention on Human Rights

If you can't explain it to an angry customer, you haven't achieved explainability

Principle 4: Robustness, Security & Safety

Case Study: Post Office Horizon System

✓ Certified robust ✓ Passed all testing ✓ Handled millions of transactions

BUT: A network bug duplicated or lost transactions

900 lives destroyed through false prosecutions

KEY LESSON: **Robustness isn't perfect testing - it's understanding failure modes**

Principle 5: Accountability

The Robodebt 'Diffusion of Responsibility'

Royal Commission: 'Diffusion of responsibility enabled the disaster'

TEST: 'Who can stop this system today if it's causing harm?'

Principle 6: Human Oversight

The Rubber-Stamping Reality

60 cases per hour = **1** minute per life-changing decision

They weren't reviewing; they were rubber-stamping

REAL OVERSIGHT REQUIRES: **Time to think • Context to understand • Authority to override**

Principle 7: Privacy & Data Governance

Case Study: Clearview AI

Scraped billions of facial images from social media

Had privacy policies ✓ Had data governance frameworks ✓

RESULT

Lawsuits & bans across multiple jurisdictions

FINDING

Violated privacy laws despite policies

The 'Creepy Test' matters more than the Compliance Test

Principle 8: Sustainable Development

The Hidden Cost of Optimisation

THE PROMISE

15% emission reduction through AI optimisation

THE REALITY

Developing region suppliers 'algorithmically invisible'

International Labour Organisation & UNCTAD Findings:

- Rural India suppliers can't meet digital data requirements
- African suppliers excluded due to 'data scarcity'
- **Developing countries bear environmental costs, receive limited benefits**

Sustainability isn't just environmental - it's about sustaining communities and dignity

Five Practices That Actually Work

1. **Test with Dirty Data**
   Use messy, incomplete, biased data - the data you'll actually get
2. **Create 'Chaos Days'**
   Deliberately break things - test contradictory inputs and edge cases
3. **Shadow Mode First**
   Run AI parallel to humans for months before going live
4. **Explanation Logging**
   Log key decision factors in plain language, not just outcomes
5. **'Stop Work' Authority**
   Someone must have power to shut down the AI - test this quarterly

Tomorrow Morning's Questions

Look at the AI system you're building or buying and ask:

1. **What biases are hiding in my training data?**
2. **Can I explain this decision to someone whose life it affects?**
3. **Who exactly can shut this down if it goes wrong?**
4. **What happens when this fails in ways I haven't imagined?**

Remember

Perfect maths can create imperfect justice

Jenny Miller's algorithm was mathematically correct.
The Horizon system worked 99% of the time.

That 1% destroyed 1,169 lives.

The ghost in the machine isn't artificial intelligence.

It's human assumption.

Key Takeaways

1 **AI Governance Must Be Proactive, Not Performative**

Implement measurable conformance mechanisms, not vague commitments

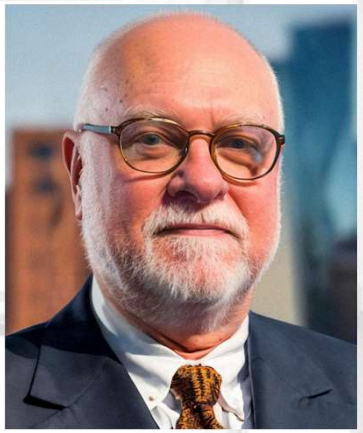
2 **Leadership Accountability Is Increasing**

Directors face growing personal legal exposure for oversight failures

3 **Global Alignment Is Crucial but Complex**

Harmonised principles emerging, but local obligations must be embedded

Thank you for attending!



darryl.carlton@me.com

Get my books at

https://technicspub.com

Twitter: @InfoGovWorld

LinkedIn: @InfoGov World Magazine

www.InfoGovWorldConference.com

AIWorldconference.ai