Dr Heinz Herrmann
Email: heinz.herrmann@torrens.edu.au
www.linkedin.com/in/heinzherrmann/

AI in Responsible Management Education & Research (RMER)
Agenda for Responsible AI (RAI) in higher education

• Definitions
  → What is/isn’t AI?
  → What is RAI?

• Using AI in L&T
  → Fight, flight or adapt? (in Q&A)
  → Job readiness

• Using RAI in responsible management research
  → Validity & reliability of AI-augmented analysis (in Q&A)
  → Role-specific managerial implications (in Q&A)
What is AI?

Artificial intelligence

Knowledge-based (since 1950s)

Metaheuristics (prominence since 1970s) or DS (2000s)

Metaheuristics & Data science

Deep learning

Breakthrough in neural-based NLP (1990s), image classification (2010s), generative (2020s)
Machine learning discovers predictions from historical data
(Herrmann & Masawi, 2022)
Generative AI creates content, disrupting “AIEd”/”EdTech”

• **Images**: DALL-E, Midjourney, Imagen, Dream, Muse AI ...

• **LLMs**: GPT, BERT, PaLM, Llama, Titan, BLOOM ...

• **Code**: Codex, Copilot, StarCoder, CodeT5 ...

• **Audio**: VALL-E, resemble.ai, AudioCraft ...

• **Multimodal**: Gato (play video games, caption images, chat, and stack blocks with a real robot arm) … → but this is still far away from AGI

• **Scientific research**: Kahubi, AvidNote, Elicit, ATLAS.ti ...

• **Malicious/illegal activities**: WormGPT, FraudGPT ...
Unified ethical principles for responsible AI (RAI) (Floridi & Cowls, 2019)

• **Beneficence**
  • Promotes human well-being and facilitates the UN SDGs

• **Non-malevolence**
  • Avoids causing harm to people and respects privacy

• **Autonomy**
  • Human autonomy over AI autonomy → “AI alignment with human goals”

• **Justice**
  • Use AI to rectify inequalities and prevent bias

• **Explicability**
  • Facilitate accountability around AI by making it more interpretable by humans
Beneficence: AI and SDGs
(Nasir et al, 2023; Vinuesa et al, 2020)

Based on curricula, declared frameworks and research papers:

• AI can enable 134 targets
• May inhibit 59 of 169 targets
• Bias towards positive outcomes?
Non-Malevolence & Justice: “AI, Algorithmic, and Automation Incidents and Controversies” (AIAAIC)

• Live spreadsheet with AI harms since 2012 → Currently 1,107 cases
  • [https://www.aiaaic.org/aiaaic-repository](https://www.aiaaic.org/aiaaic-repository)

• Not captured in that database are unreported cases or future implications:
  • Structural unemployment
  • A digital version of Taylorism
  • Military use
  • Machine consciousness
  • Super-intelligence → the Autonomy principle
Managerial practice:
AI will cause a job polarisation in the form of a “dumbbell shape”

By 2033, AI may generate more than half a billion net-new human jobs (Gartner, 2023), but this poses a challenge for SDG 4 (Education)
When research articles deal with “implications for managerial practice”

• It should address “the degree to which a specific manager in an organization perceives academic knowledge to aid his or her job-related thoughts or actions in the pursuit of organizational goals” (Jaworski, 2011)

• PRME’s Partnership principle → role-relevant research

• Guiding questions
  1. What is the target manager’s job role? (CEO, HR Director, …)
  2. Which role task is the focus of the research? (Compliance, recruitment, …)
  3. When is the impact to occur? (Now/future)
  4. What is the desired impact? (Thinking/action)
  5. Which research part will achieve this impact? (Findings, framework, …)
Appendix
for Q&A
and references
The Arcanum of AI
(Herrmann, 2022)
Mapping of RAI and PRME Principles

- Beneficence
- Non-malevolence
- Justice
- Autonomy
- Explicability

- Purpose
- Values
- Method
- Research
- Partnership
- Dialogue
Beneficence: Based on industry

• Currently, AI development focuses on economic growth, neglecting important societal & environmental issues → Frequent alibi-driven “ethics washing” in published corporate RAI frameworks

• “Unless we act now, the 2030 Agenda will become an epitaph for a world that might have been.” (Antonio Guterres, 2023)

• **The problem:** Few industry leaders outside B Corps or Social Enterprises leaders apply systems thinking in their decision-making towards sustainability

• **A step forward:** Understanding corporate politics is increasingly important (that is the most popular part of Stamford’s curriculum, *The Economist*, 2023)

• **More radically:** How can higher education assist in the redesign of the globally different flavours of capitalism towards SDG goals?
Non-Malevolence, Justice & Explicability: Regulation & Standards

- 123 AI-related bills passed into law since 2016 in 127 countries (HAI, 2023)
- European AI Act is expected to pass this year
- US principles currently are voluntary and rely on self-regulation by companies
- Voluntary commitment from OpenAI, Amazon, Anthropic, Google, Inflection, Meta and Microsoft to US regulation of “frontier models” (> currently released GAI)
- The EU-US Trade and Technology Council develops a common understanding of trustworthy AI and works collaboratively on international AI standards
- IEEE 7000 standard for ethics in the design of AI
The European Commission’s High-Level Expert Group’s Ethics Guidelines for Trustworthy AI (2019)

“Ethical reflection ... can stimulate new kinds of innovations that seek to foster ethical values, such as those helping to achieve the UN Sustainable Development Goals, which are firmly embedded in the ... EU Agenda 2030.”
A European RAI perspective:
RAI = ethical principles + their governance
(Macnaghten et al, 2014)

- Anticipation
- Inclusion
- Reflectiveness
- Responsiveness

35% of RAI publications in Scopus relate to governance of AI. Note that Governments also govern with AI.
A global perspective from current policies
(Wittrock et al, 2021)

• Ethics
• Gender equality and diversity
• Open access and open science
• Science education
• Public engagement
What do internal review boards need to consider for research? (Herrmann & Cameron, 2023)

**Ethics Committee**

**Ethical Principles:**
- Explicability/Transparency
- Beneficence
- Non-Malevolence
- Justice
- Autonomy

**Government Policy:**
- Ethics, Gender Equality & Diversity
- Open Access & Open Science
- Science Education
- Public Engagement

**RAI Governance Process Dimensions**

- Diversity
- Reflexivity
- Openness
- Transparency
- Anticipation, Inclusion, Responsiveness, Adaptation

**AI Product**

- AI Engineers Create Value-Laden Product
- Users of AI Product
- Public Stakeholders

**Interpretability:** Understand how an AI tool works & how to debug problems

**Explicability:** Understand why & why not predictions were made
We need to prepare students for their future: modes of managerial decision-making need to be taught (Ross & Taylor, 2021)

- **Humans in the loop (HITL)**
  - Humans make the decision and AI provides decision support only

- **Humans in the loop for exceptions (HITLFE)**
  - Humans handle exceptions only

- **Human on the loop (HOTL)**
  - Humans review the decision outcomes and adjust AI parameters for future decisions

- **Human out of the loop (HOOTL)**
  - Humans intervene only by setting new constraints and objectives. AI parameter adjustments are automated, based on feedback from humans
  - By 2030, such automated decisions are predicted to cause USD100 billion in losses from asset damage (Gartner, 2023) → Reinforce the non-malevolence principle
In research: The *cited* half-life of knowledge (Davis & Cochran, 2015)

Measures the **longevity of citations** in terms of the median age of citations → examples:

- **Business & management:** 9 years
- **Education & educational research:** 8 years
- **Artificial intelligence:** 7 years
- **Robotics:** 6 years
From research to the half-life of professional expertise

• Measures the **obsolescence of knowledge** in terms of the time for half of the knowledge to become superseded

• My doctorate (2001) used Data Science, based on statistical modelling → a decade later the field shifted to AI-based machine learning

• **General half-life examples** (Germain, 2021):
  • 1990s: 10—15 years
  • 2020—2025: 3—5 years

• **Lifelong learning** is ever so important → JIT learning/micro-credentialling

• **Threshold learning** releases dopamine in the brain, which is highly addictive as it “rewires” synaptic connections between neurons
“AIEd”/”EdTech” in higher education until 2022 focused on efficiency (Crompton & Burke, 2023)

- **Continuous improvement** → learning analytics
- **Assessment evaluation** → more reliable for feedback on student drafts
- **Predicting performance**, including at-risk students
- **AI Assistant**, including student experience → chatbots and recommendations
- **Intelligent Tutoring System (ITS)** → personalised learning for L&T effectiveness
  - Over-reliance on personalisation/hyper-personalisation has problems → “Learning how to learn” (academic literacy) and student adaptivity need to be key learning objectives
  - Personalisation on a learner’s knowledge or affective states is effective, but science is clear that not all personalisation is effective → learning styles or attention span
Risks with ChatGPT (= GPT foundation + finetuned with RLHF + chatbot) and kin

• **Plagiarism**
  • AI content detectors are unreliable to date with a high false positive rate → use them as a flag only

• **Watermarking** (embedding a statistical pattern into word/punctuation choices for **LLMs**, or pixels for **images** in their “latent space”) → alternatively, **authentic assessments**

• **Bias** → refer back to the **Justice** principle in RAI

• **Hallucinations**, incl. fake references → **Wolfram plugins**

• **Privacy** → Don’t upload personally identifiable information when using LLMs for feedback on student work
Ideas for L&T with generative AI

• **Students’ use of generative AI**
  - Reader’s and Writer’s block → overcome procrastination
  - Practice academic literacy → gen AI as an extra teacher
  - If none of the above → plagiarism/cheating

• **Scaffold the use of AI tools like other tools are taught** → ask students to critique AI-produced content
  - Focus shifts from fundamental (explicit & codifiable/programmatic) knowledge and mechanical skills towards creativity & critical thinking that is integrated with tacit knowledge (intuition/System 1) → compare AI-produced content with reliable, valid sources of information

• **Show & tell**: feed the assessment brief into AI
  - Discuss the results with students and whether the subject material was covered
  - Demonstrate hallucinations and fake references
Generative AI in L&T: UNESCO (SDG 4) start-up guide for ChatGPT (Sabzalieva & Valentini 2023)

<table>
<thead>
<tr>
<th>Role</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Possibility engine</td>
<td>AI generates alternative ways of expressing an idea</td>
</tr>
<tr>
<td>Socratic opponent</td>
<td>AI acts as an opponent to develop and argument</td>
</tr>
<tr>
<td>Collaboration coach</td>
<td>AI helps groups to research and solve problems together</td>
</tr>
<tr>
<td>Guide on the side</td>
<td>AI acts as a guide to navigate physical and conceptual spaces</td>
</tr>
<tr>
<td>Personal tutor</td>
<td>AI tutors each student and gives immediate feedback on progress</td>
</tr>
<tr>
<td>Co-designer</td>
<td>AI assists throughout the design process</td>
</tr>
<tr>
<td>Exploratorium</td>
<td>AI provides tools to play with, explore and interpret data</td>
</tr>
<tr>
<td>Study buddy</td>
<td>AI helps the student reflect on learning material</td>
</tr>
<tr>
<td>Motivator</td>
<td>AI offers games and challenges to extend learning</td>
</tr>
<tr>
<td>Dynamic assessor</td>
<td>AI provides educators with a profile of each student’s current knowledge</td>
</tr>
</tbody>
</table>
Generative AI in management research

- Create abstracts
- Critique before submission to journal
- Get opposite viewpoints
- Identify research gaps
- Systematic reviews generate new hypotheses
- Suggest research method
- Interpret data
- Generate interview/survey questions
- Code interviews/thematic analysis
Explicability: AI in research

• Cloud software can threaten replicability of research

• GPT-4 & GPT-3.5’s reasoning performance changed March – June 2023 (Chen et al, 2023)

• Commercial LLMs avoided peer review processes

• Closed & proprietary models make them unfit for responsible use in research

• Other purported “open” models often involve undocumented data of dubious legality, few share the instruction tuning and scientific documentation (Liesenfeld et al, 2023)
References


• Herrmann, H., & Masawi, B. (2022). Three and a half decades of artificial intelligence in banking, financial services, and insurance: A systematic evolutionary review. *Strategic Change, 31*(6), 549-569. [https://doi.org/10.1002/jsc.2525](https://doi.org/10.1002/jsc.2525)
References

• Jaworski, B. (2011). On managerial relevance. *Journal of Marketing, 75*(4), 211-224. [https://doi.org/https://doi.org/10.1509/jmkg.75.4.211](https://doi.org/https://doi.org/10.1509/jmkg.75.4.211)


