Recognise, Evaluate and Address The Downsides of GenAI

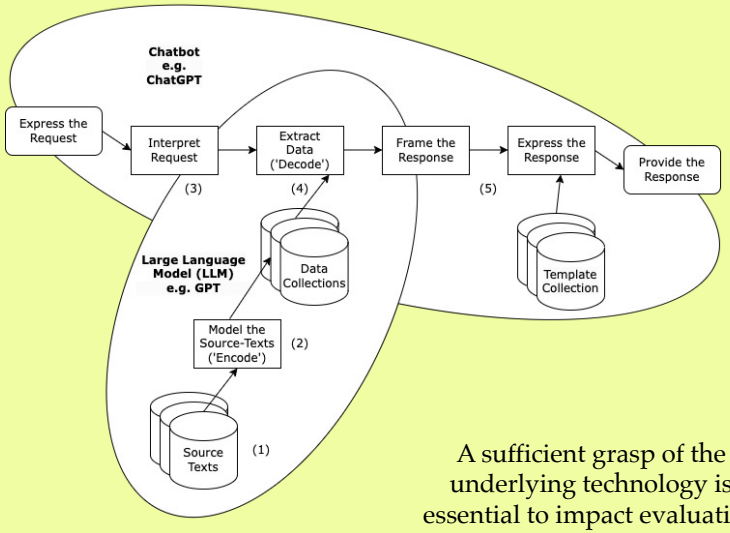
Roger Clarke
Xamax Consultancy, Canberra
Visiting Professor, UNSW Law and ANU RSCS

http://rogerclarke.com/EC/GAIH-DHR.pdf (Slides)
http://rogerclarke.com/EC/RGAI-C.html (PrePrint)

Digital Health Research CoP
14 March 2025

Copyright 2018-25 XAMAX Consultancy
SOME RIGHTS RESERVED
FREE FOR EDUCATION www.aesharenet.com.au/FfE2/

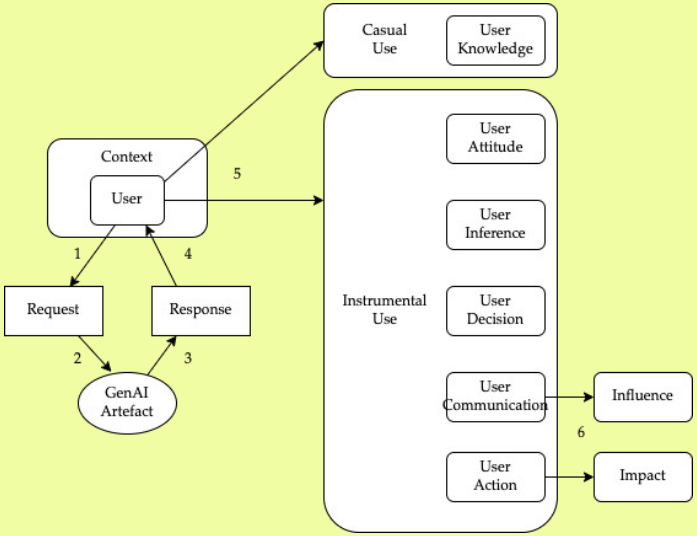
A Process View of GenAI Artefacts



A sufficient grasp of the underlying technology is essential to impact evaluation

Copyright 2018-25 XAMAX Consultancy

GenAI Categories of Use



Copyright 2018-25 XAMAX Consultancy

A Typology of Requests / Use Cases

- **Question / Information-Request**
  What brandnames is this generic equivalent to?
- **Composition**
  Summarise for me
  What are the latest findings on ...?
  Generate me some synthetic data
- **Dialogue / Interrogation / What-If Analysis**
  A Cumulative Series of Questions

Copyright 2018-25 XAMAX Consultancy

A Typology of Requests / Use Cases

- Question, Composition or Dialogue
- **About Codified Knowledge** (cf. refined free-text search):
  - Exposition
  - Description
  - Explanation
- **About a Case**:
  - Description
  - Explanation
  - Prediction

A Typology of Requests / Use Cases

- Question, Composition or Dialogue
- About Codified Knowledge (cf. refined free-text search):
  - **Exposition** Physiology, Disease, Condition
  - **Description** Symptoms, Contra-Indications
  - **Explanation** Determinants, Environmental Factors
- About a Case:
  - **Description** Case Notes, Discharge Summary
  - **Explanation** Exercise Routine, Course of Treatment
  - **Prediction** Prognosis

Potentially Harmful Attributes of GenAI (20)

- Attributes of Source-Text Collections (7)
- Limited Understanding of Language Structure (2)
- Absence of any Real-World Referents (5)
- Attributes of the Generative Function (6)

1. Attributes of Source-Text Collections

1. Indiscriminate Acquisition of vast corpora rather than Curation of Source-Texts
2. Inadequate Content Quality-Assurance
3. Embedded Misinformation and Disinformation

6. Unintended Effects of 'Scrubbing'

==>> Use only Custom GenAI Artefacts
==>> Use scrubbing only very carefully
==>> Beware of generic Foundation Models

2. Limited Understanding of Language Structure

1. Linguistic Analysis dominated by Word Sequence with limited reflection of full linguistic grammars
2. Linguistic Synthesis dominated by Word Sequence Probabilities

==>> 'It's merely a Large Limited-Syntax Model'

==>> 'Be aware that it's a Stochastic Parrot'

==>> 'Beware of synthetic authority'

==>> 'Beware of human gullibility and laziness when looking at apparently authentic text' ('Remember Eliza!')

Copyright 2018-25 XAMAX Consultancy

3. Absence of any Real-World Referents

Absence of Understanding of:

- Content, Context
- Requestor / Audience
- Other Stakeholders
- Consequences for Stakeholders

Absence of any Capacity for 'Common Sense'

Minimal Syntactics, No Semantics or Pragmatics:

- Syntactics relations ... within data
- **Semantics ... between data and real-world things**
- **Pragmatics ... between data and people**

==>> Appreciate Limitations, Apply Scepticism.
People and Organisations are Responsible for Emotional Intelligence, Judgements, Decisions

Copyright 2018-25 XAMAX Consultancy

4. Attributes of the Generative Function

2. Limited **Explainability** of the Rationale Underlying Responses

6. **Artefacts / Hallucinations** from Source Encoding / Decoding

GenAI generally:

- is not Rational
- may be Irrational
- is in all cases **at least A-Rational**

Copyright 2018-25 XAMAX Consultancy

4. Attributes of the Generative Function

2. Limited **Explainability** of the Rationale Underlying Responses
3. **Limited Capability to Provide Source Quotations, Citations**
6. **Artefacts / Hallucinations** from Source Encoding / Decoding

==>> Unless Auditability and Accountability are optional:

- A-Rationality is a Contraindication for GenAI Use
- Believe in XAI only if it moves beyond aspiration and actually delivers the goods

Copyright 2018-25 XAMAX Consultancy

Accountability depends on Transparency of Rationale

- **A-rationality**
  - Unexplainability
    - Unreplicability
      - Unauditability
        - Uncorrectability
          - **Unaccountability**

A Test-Case: Diagnosis

A process to classify a condition experienced by a patient

- Access to Case Info (Signs, Symptoms, Circumstances)
- Access to Scientific Information, 'reliable' and 'up-to-date'
- Identification of Sci-Info 'relevant' to the Case Info
- Summarise Case Info and Sci Info, identify Options
- Offer a Provisional Diagnosis, or a Differential Diagnosis plus Case Info needed to distinguish between alternatives
- 'Participate' in a Dialogue / Interrogation
- 'Recalibrate' based on new Case Information

Recognise, Evaluate and Address The Downsides of GenAI

Roger Clarke
Xamax Consultancy, Canberra
Visiting Professor, UNSW Law and ANU RSCS

http://rogerclarke.com/EC/GAIH.pdf (Slides)
http://rogerclarke.com/EC/ RGAI-C.html (PrePrint)

Digital Health Research CoP
14 March 2025

SOME RIGHTS RESERVED

FREE FOR EDUCATION www.aesharenet.com.au/FfE2/

Copyright 2018-25 XAMAX Consultancy

Autonomous GenAI Artefacts or Complementary Artefact Intelligence?

		Function of the Artefact	Function of the Human
	0	~~NIL~~	**Analyse, Decide, Act**
Decision Support System	1	Analyse Options	**Analyse, Decide, Act**
	2	Advise re Options	**Analyse, Decide, Act**
	3	Recommend Act	**Analyse, Approve/Reject Act**
Decision System	4	**Notify Impending Act**	Override/Veto Impending Act
	5	**Act and Inform**	Interrupt/Suspend/Cancel an Act
	6	**Act**	~~NIL~~

Copyright 2018-25 XAMAX Consultancy

'Why the World Wants Controls over Artificial Intelligence'
Computer Law & Security Review 35, 4 (2019) 423-433
https://rogerclarke.com/EC/AII.html#TAA

10 Groups of 50 Principles for Responsible AI

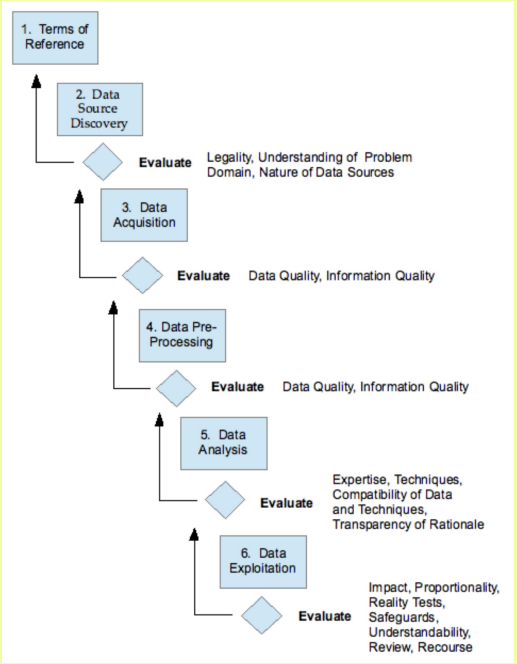
1. Evaluate Positive and Negative Impacts
2. Complement Humans
3. Ensure Human Control
4. Ensure Human Wellbeing and Safety
5. Ensure Consistency with Human Values and Human Rights
6. Deliver Transparency and Auditability
7. Embed Quality Assurance
8. Exhibit Robustness and Resilience
9. Ensure Accountability for Legal and Moral Obligations
10. Enforce, and Accept Enforcement of, Liabilities and Sanctions

Copyright 2018-25 XAMAX Consultancy

'Principles and Business Processes for Responsible AI'
Computer Law & Security Review 35, 4 (2019) 410-422
http://rogerclarke.com/EC/PRAI.html#Princ

A Business Process for Responsible AI/ML

http://rogerclarke.com/EC/BDBP{.html, .pdf}
Clarke & Taylor (2018)



Copyright 2018-25 XAMAX Consultancy

AI embodies errors of inference, decision and action
arising from the independent operation of artefacts,
for which **no rational explanation is available**,
which results in inferences, decisions and actions
incapable of investigation, correction and reparation

A Summary of the Sources of AI's Threats

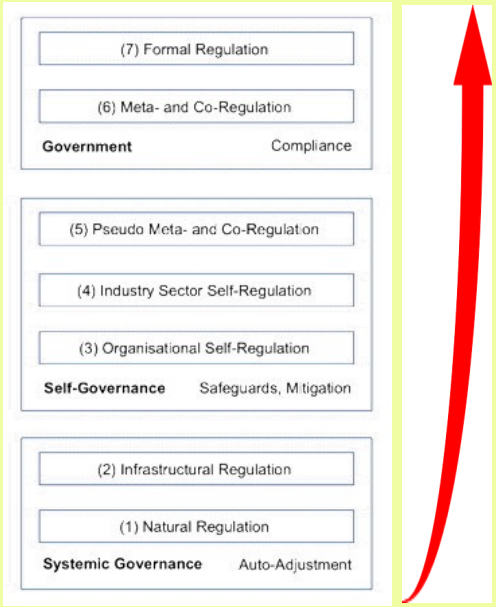
1. Artefact **Autonomy**
2. **Inappropriate Assumptions** ... about Data
3. ... and about the Inferencing Process
4. **Opaqueness** of the Inferencing Process
5. **Irresponsibility**

Socio-Political Impacts and Implications

- ***De Facto* Delegation**
  'The computer says no'
- **Unexplainability**
  Accountability Undermined
- **Unfair Decisions, Actions**
  Discriminatory Behaviour
- **Economic, Social Scoring**
  Non-Conformist Victimisation
- **Undefendable Accusations**
  Power, Information Asymmetry
- **'Predestination'**
  e.g. Predictive Policing
- **People-Replacement**
  Effect on Income Distribution
- **Denial of Services, of Movement, of Identity**
  Public Resentment, Violence

Proposition:

A Hierarchy of Regulatory Mechanisms



Formalism
Compulsion

50 Principles for Responsible AI

1. Evaluate Positive and Negative Impacts

1.1 Conceive and design only after ensuring adequate **understanding** of purposes and contexts
1.2 **Justify objectives**
1.3 Demonstrate the achievability of postulated benefits **('justification')**
1.4 Conduct **impact assessment**
1.5 Publish **sufficient information to stakeholders** to enable them to conduct impact assessment
1.6 Conduct **consultation with stakeholders** and enable their participation
1.7 **Reflect stakeholders' justified concerns**
1.8 Justify negative impacts on individuals ('**proportionality**')
1.9 Consider less harmful ways of achieving the same objectives ('**mitigation**')