Responsible AI Technologies, Artefacts, Systems and Applications
50 Principles

The following Principles are intended to be applied by the entities responsible for all phases of AI research, invention, innovation, dissemination and application. They were derived by consolidating elements from 30 international sources on 'Ethical Analysis and IT' and 'Principles for AI'.

1. Assess Positive and Negative Impacts and Implications

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
1.4 Conduct impact assessment, including risk assessment from all stakeholders' perspectives
1.5 Publish sufficient information to stakeholders to enable them to conduct impact assessment
1.6 Conduct consultation with stakeholders and enable their participation in design
1.7 Reflect stakeholders' justified concerns in the design
1.8 Justify negative impacts on individuals ('proportionality')
1.9 Consider alternative, less harmful ways of achieving the same objectives

2. Complement Humans

2.1 Design as an aid, for augmentation, collaboration and inter-operability
2.2 Avoid design for replacement of people by independent artefacts or systems, except in circumstances in which those artefacts or systems are demonstrably more capable than people, and even then ensuring that the result is complementary to human capabilities

3. Ensure Human Control

3.1 Ensure human control over AI-based technology, artefacts and systems
3.2 In particular, ensure control over autonomous behaviour of AI-based technology, artefacts and systems
3.3 Respect people's expectations in relation to personal data protections, including their awareness of data-usage, their consent, data minimisation, public visibility and design consultation and participation, and the relationship between data-usage and the data's original purpose
3.4 Respect each person's autonomy, freedom of choice and right to self-determination
3.5 Ensure human review of inferences and decisions prior to action being taken
3.6 Avoid deception of humans
3.7 Avoid services being conditional on the acceptance of AI-based artefacts and systems

4. Ensure Human Safety and Wellbeing

4.1 Ensure people's physical health and safety ('nonmaleficence')
4.2 Ensure people's psychological safety, by avoiding negative effects on their mental health, emotional state, inclusion in society, worth, and standing in comparison with other people
4.3 Ensure people's wellbeing ('beneficence')
4.4 Implement safeguards to avoid, prevent and mitigate negative impacts and implications
4.5 Avoid violation of trust
4.6 Avoid the manipulation of vulnerable people, e.g. by taking advantage of individuals' tendencies to addictions such as gambling, and to letting pleasure overrule rationality

5. Ensure Consistency with Human Values and Human Rights

5.1 Be just / fair / impartial, treat individuals equally, and avoid unfair discrimination and bias, not only where they are illegal, but also where they are materially inconsistent with public expectations
5.2 Ensure compliance with human rights laws
5.3 Avoid restrictions on, and promote, people's freedom of movement
5.4 Avoid interference with, and promote privacy, family, home or reputation

5.5 Avoid interference with, and promote, the rights of freedom of information, opinion and expression, of freedom of assembly, of freedom of association, of freedom to participate in public affairs, and of freedom to access public services

5.6 Where interference with human values or human rights is outweighed by other factors, ensure that the interference is no greater than is justified ('harm minimisation')

6. Deliver Transparency and Auditability

6.1 Ensure that the fact that a process is AI-based is transparent to all stakeholders

6.2 Ensure that data provenance, and the means whereby inferences are drawn from it, decisions are made, and actions are taken, are logged and can be reconstructed

6.3 Ensure that people are aware of inferences, decisions and actions that affect them, and have access to humanly-understandable explanations of how they came about

7. Embed Quality Assurance

7.1 Ensure effective, efficient and adaptive performance of intended functions

7.2 Ensure data quality and data relevance

7.3 Justify the use of data, commensurate with each data-item's sensitivity

7.4 Ensure security safeguards against inappropriate data access, modification and deletion, commensurate with its sensitivity

7.5 Deal fairly with people ('faithfulness', 'fidelity')

7.6 Ensure that inferences are not drawn from data using invalid or unvalidated techniques

7.7 Test result validity, and address the problems that are detected

7.8 Impose controls in order to ensure that the safeguards are in place and effective

7.9 Conduct audits of safeguards and controls

8. Exhibit Robustness and Resilience

8.1 Deliver and sustain appropriate security safeguards against the risk of compromise of intended functions arising from both passive threats and active attacks, commensurate with the significance of the benefits and the potential to cause harm

8.2 Deliver and sustain appropriate security safeguards against the risk of inappropriate data access, modification and deletion, arising from both passive threats and active attacks, commensurate with the data's sensitivity

8.3 Conduct audits of the justification, the proportionality, the transparency, and the harm avoidance, prevention and mitigation measures and controls

8.4 Ensure resilience, in the sense of prompt and effective recovery from incidents

9. Ensure Accountability for Legal and Moral Obligations

9.1 Ensure that the responsible entity is apparent or can be readily discovered by any party

9.2 Ensure that effective remedies exist, in the form of complaints processes, appeals processes, and redress where harmful errors have occurred

10. Enforce, and Accept Enforcement of, Liabilities and Sanctions

10.1 Ensure that complaints, appeals and redress processes operate effectively

10.2 Comply with external complaints, appeals and redress processes and outcomes, including, in particular, provision of timely, accurate and complete information relevant to cases

---

Published Versions

Clarke R. (2019) 'Principles and Business Processes for Responsible AI' Computer Law & Security Review 35, 4 (2019) 410-422, PrePrint at http://www.rogerclarke.com/EC/AIP.html

Clarke R. (2020) 'Principles for Responsible AI' Xamax Consultancy Pty Ltd, January 2020, at http://www.rogerclarke.com/EC/RAIC.html

---