UCL

MEASURING AI ETHICS

Authors

Paul Walton (Capgemini)
Jasmin Booth (Capgemini)
Jamie Rich (Capgemini)
Dino Mariutti (UCL MBA student)
Tom Weston (UCL MBA student)

UCL
SCHOOL OF
MANAGEMENT

ANALYTICS
LAB

Capgemini

ABSTRACT

IMPLEMENTING ARTIFICIAL INTELLIGENCE (AI) ETHICALLY IS ESSENTIAL, AND UNDERSTANDING WHETHER ANY IMPLEMENTATION IS ETHICAL REQUIRES MEASUREMENT.

But measurement is challenging because:

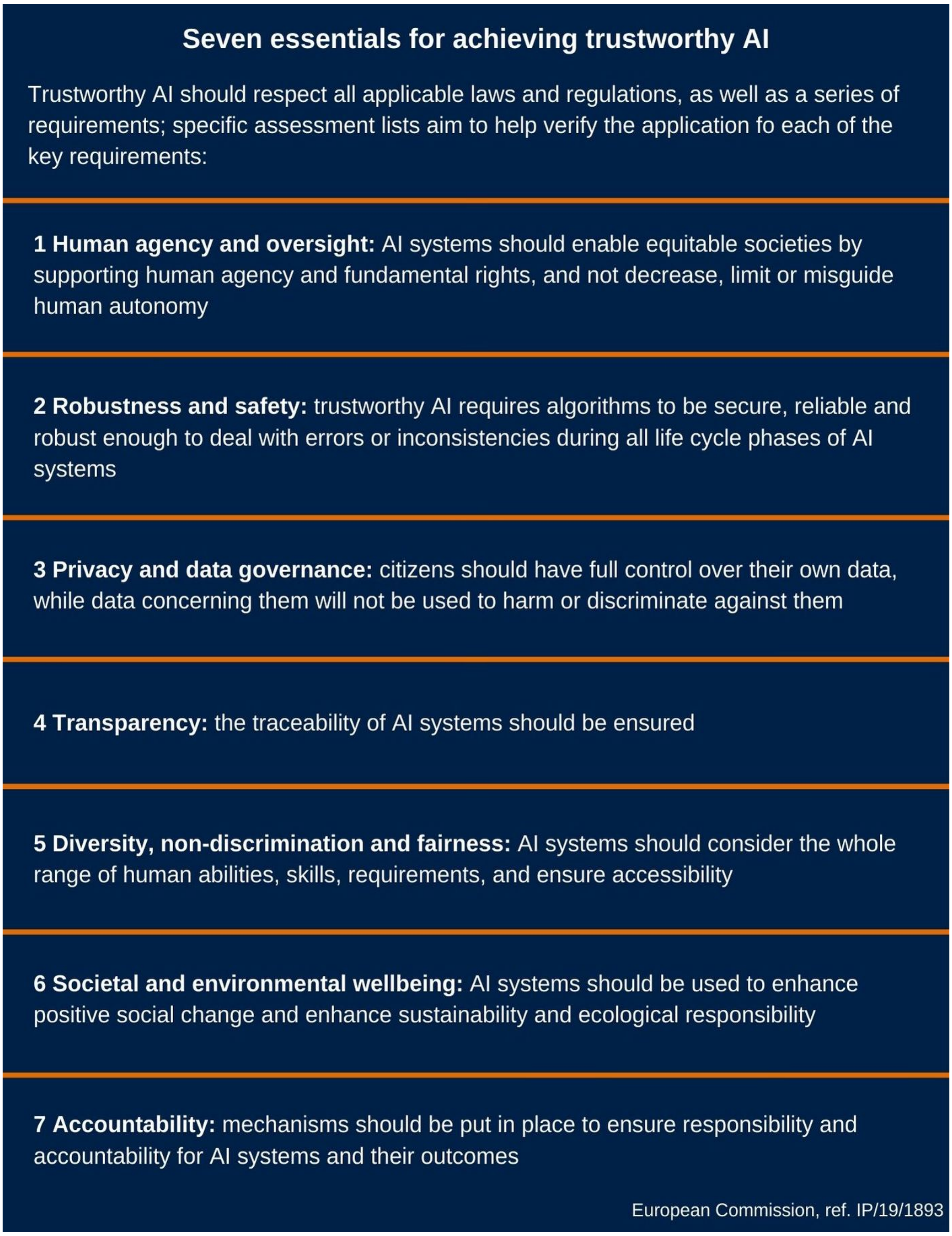
- Understanding AI ethics requires human judgement because measures will only provide an incomplete picture;
- Human judgement is not consistent and cannot guarantee objectivity;
- The root causes of ethical issues may be hard to measure but measuring only what is easy will cause ethical risks;
- Automation using AI may accelerate rare and damaging ("Black Swan") outcomes that any measurement process may not be reliable enough to identify.

This article explores why an ethical approach is required and what needs to be measured. It concludes by examining the challenges posed by measuring AI ethics.

AN ETHICAL APPROACH TO AI

The need for an ethical approach to AI is clearly established. The use of AI is growing fast and it is becoming embedded in multiple forms in business processes. At the same time, it is available to all developers and even (through "no/low code" technology) to business users. Without a clear approach to AI ethics, AI will be in the hands of business users and technical teams without a clear understanding of the risks or the means of mitigating them.

The principles of AI ethics are expressed differently by different authors but there is agreement to the underlying ideas. For example, the EU has published Ethical Guidelines for Trustworthy AI. According to the guidelines, trustworthy AI should be lawful, ethical and robust. The ethical component is captured in the principles shown in Figure 1.



Seven essentials for achieving trustworthy AI

Trustworthy AI should respect all applicable laws and regulations, as well as a series of requirements; specific assessment lists aim to help verify the application fo each of the key requirements:

1 Human agency and oversight: AI systems should enable equitable societies by supporting human agency and fundamental rights, and not decrease, limit or misguide human autonomy

2 Robustness and safety: trustworthy AI requires algorithms to be secure, reliable and robust enough to deal with errors or inconsistencies during all life cycle phases of AI systems

3 Privacy and data governance: citizens should have full control over their own data, while data concerning them will not be used to harm or discriminate against them

4 Transparency: the traceability of AI systems should be ensured

5 Diversity, non-discrimination and fairness: AI systems should consider the whole range of human abilities, skills, requirements, and ensure accessibility

6 Societal and environmental wellbeing: AI systems should be used to enhance positive social change and enhance sustainability and ecological responsibility

7 Accountability: mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes

European Commission, ref. IP/19/1893

Figure 1: EU principles for ethical AI

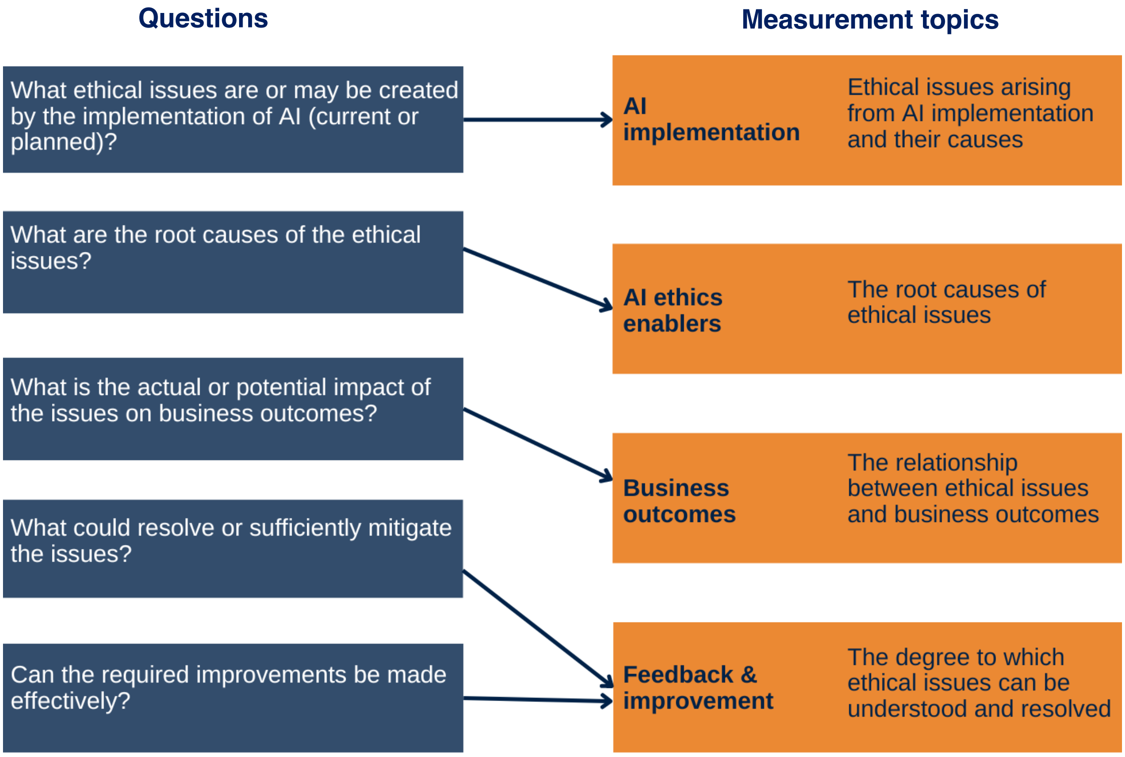
But the principles are not, in themselves, enough. Organisations need to understand how well the principles are being adopted – they need measures, but what do they need to measure?

MEASURES OF AI ETHICS

The journey to the widespread implementation of AI is just starting in many organisations so implementing an ethical approach to AI is necessarily iterative. Organisations need to be able to answer the following questions:

- What ethical issues are or may be created by the implementation of AI (current or planned)?
- What are the root causes of the ethical issues?
- What is the actual or potential impact of the issues on business outcomes?
- What could resolve or sufficiently mitigate the issues?
- Can the required improvements be made effectively?

Answering these questions needs measures covering the topics shown in Figure 2.



Questions

Measurement topics

What ethical issues are or may be created by the implementation of AI (current or planned)?

What are the root causes of the ethical issues?

What is the actual or potential impact of the issues on business outcomes?

What could resolve or sufficiently mitigate the issues?

Can the required improvements be made effectively?

AI implementation – Ethical issues arising from AI implementation and their causes

AI ethics enablers – The root causes of ethical issues

Business outcomes – The relationship between ethical issues and business outcomes

Feedback & improvement – The degree to which ethical issues can be understood and resolved

Figure 2: Mastering AI ethics transformation

Measures must accommodate all ways of implementing AI, including:

- By professional development teams and data scientists – this is the main focus of the literature;
- Through the procurement of software products or services that incorporate AI – in this case the development process may not be visible or accessible;
- Through the use of “no code/low code” tools (like the Microsoft Power Platform, for example), which provide access to AI for non-professional developers (so-called “citizen developers).

For each of these types of implementation the key principle is "ethical by design". (This general idea is well understood in other domains. For example, a "secure by design" approach has been implemented in many organisations to ensure the security of technology implementations.) The key idea is that informed consideration is given to AI ethics throughout the lifecycle so that it is built into decision-making and implementation activities rather than just being checked at the end.

However, the root causes of ethical issues may lie in elements of the operating model outside these different types of implementation. Organisational culture may inhibit some of the changes required. For example, staff may be uncertain about the impact on their jobs and unwilling to adopt AI. Procurement and commercial processes may not incorporate AI ethics. Issues with data quality and governance may reduce effectiveness. There is a complex relationship between these root causes, ethical issues and business outcomes. Understanding this relationship (for example, potential reputational damage) is critical.

Finally, even if it is clear what needs to be changed, can the changes be delivered effectively and predictably and without unexpected consequences?

The measurement approach needs to include all of these elements and is examined in more detail in the Appendix which shows the scale and complexity of the measurement required.

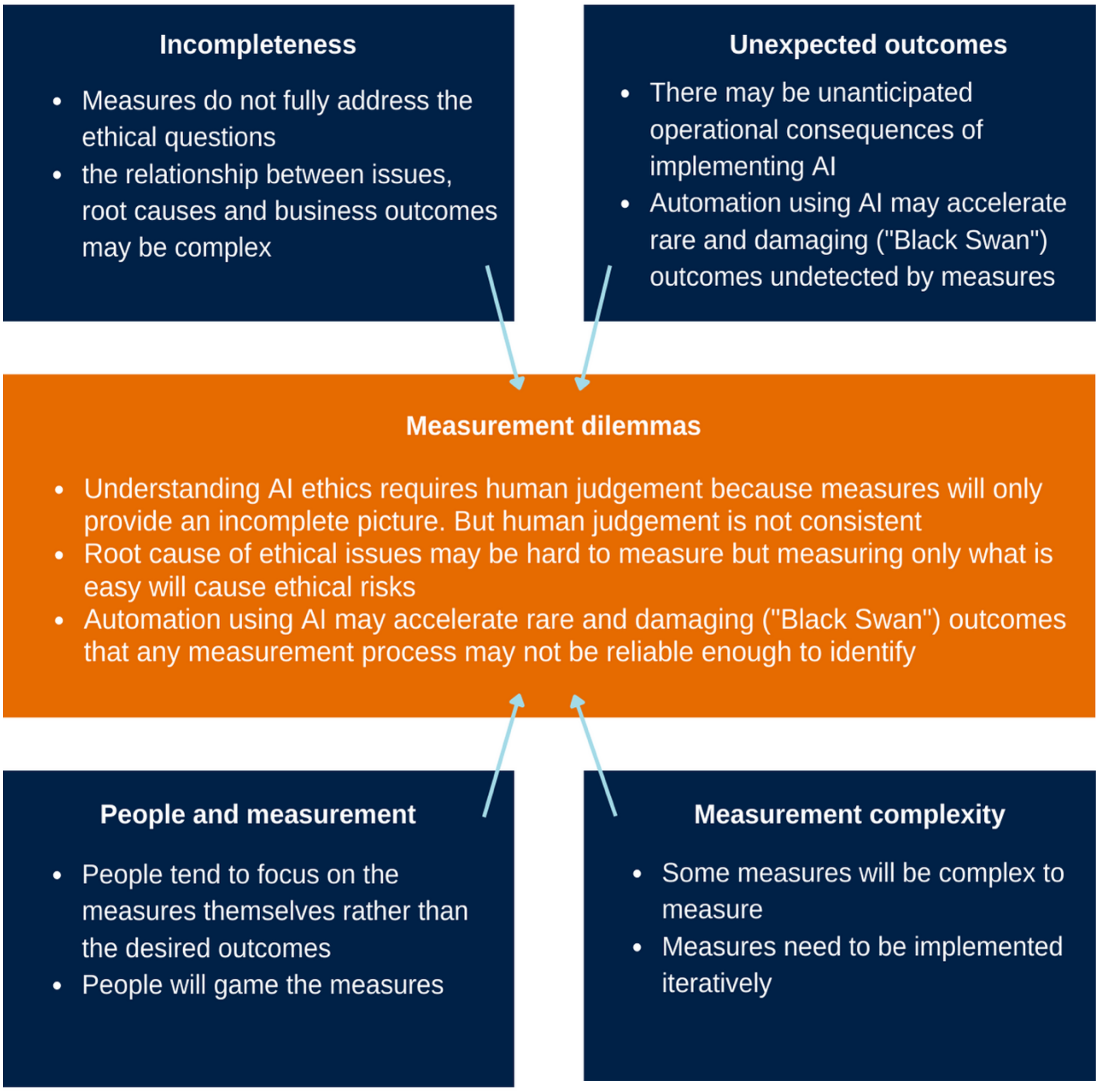
Table 1 highlights some (among many) of the theoretical and pragmatic questions and difficulties associated with the measurement of AI ethics for the motor insurance claims management process. The table highlights the difficulties of embedding a rigorous approach to AI ethics.

Table 1: Ethical issues in motor insurance claims

Process element	Potential AI role	Ethical questions and considerations
Prevent accidents	Understand driver behaviour and influence drivers to improve	• How far can this go without violating the “human agency” principle? What degree of psychological manipulation is acceptable? • Will people agree, knowingly, to how their data is used? What level of consent is needed for different uses? Does the need for consent imply any form of discrimination? • When does understanding driver behaviour violate the need for privacy? Many driver behaviours (e.g. driving at night) may not be a choice (e.g. for shift workers). • Is there sufficient evidence that the influence techniques work well enough? When is a statistical relationship sufficiently robust?
Acquire and update information about the claim	Understand the complexity and risk (including for fraud and criminal activity) associated with the incident through analysis of images, documents and interaction with participants (including witnesses, breakdown services, medical facilities, garages, lawyers, …)	Image, document and interaction analysis and their relationship with risk are all subject to bias and the impact of previous risk thresholds
Acquire and update information about the driver	Understand the propensity for poor driving behaviour, fraud or criminal activity from information about the driver and associates to compare with industry sources	• Increasing the level of profiling risks violation of privacy and fundamental rights and could lead to a form of social credit • How much data is it fair to collect (e.g. from social media) and to what extent is the collection context-specific (e.g. with respect to potential criminality)? To what extent does the transparency principle imply that the customer can see what data has been used? How much consent is or should be needed for access to this data? • Profiling requires very robust systems and robust rationale to ensure that the results are fair (how reliable is historical data about people as they change?) • Profiling may violate the diversity and fairness principle because of bias in the data
Interact with parties involved	Manage activities, communications and relationships with participants	• To what extent should AI demonstrate transparency in the interactions? How transparent should the use of data be and to what extent should it be available to challenge? • How can the AI explain decisions, as well as the use of data well enough to different parties? • What is an acceptable level of emotion sensing and psychological manipulation in interactions (e.g., Microsoft have withdrawn their emotion sensing technology)? • How much will human agency be diminished if users cannot speak to anyone (because of AI-driven interactions, extending current chatbot use)? What is the impact of the diversity principle in establishing fair interaction?
Decide the next step in the case	Profile the level of risk based on updated information about the incident, driver and other risks. Decide the process to follow and the need for specialist intervention	• What is an ethical threshold for AI (versus human) decision-making? • Can decisions be explained to customers and other stakeholders well enough? • As more decisions are made by AI, what impact will this have on human decision-making?
Litigate	Provide evidence in any claim taken to court	Are the decisions made by AI sufficiently robust and explainable? To what extent is regulation or accreditation required to provide a level playing field?
Improve	Analyse the process followed and recommend improvements	If AI is used to recommend improvements to the claim process, will those recommendations encompass ethical considerations?

MEASUREMENT CHALLENGES

This example shows that there are many challenges in measuring what's needed. The challenges are of different types, ranging from the theoretical to the pragmatic. Figure 3 provides a summary.



Incompleteness

- Measures do not fully address the ethical questions
- the relationship between issues, root causes and business outcomes may be complex

Unexpected outcomes

- There may be unanticipated operational consequences of implementing AI
- Automation using AI may accelerate rare and damaging ("Black Swan") outcomes undetected by measures

Measurement dilemmas

- Understanding AI ethics requires human judgement because measures will only provide an incomplete picture. But human judgement is not consistent
- Root cause of ethical issues may be hard to measure but measuring only what is easy will cause ethical risks
- Automation using AI may accelerate rare and damaging ("Black Swan") outcomes that any measurement process may not be reliable enough to identify

People and measurement

- People tend to focus on the measures themselves rather than the desired outcomes
- People will game the measures

Measurement complexity

- Some measures will be complex to measure
- Measures need to be implemented iteratively

Figure 3: Challenges in AI ethics measurement

Incompleteness

There is a large gap between the high-level ethical principles (which touch on long established questions in moral philosophy) and what can be measured in detail. The measurement of AI ethics is immature and is still the subject of considerable research. Some researchers even argue that it is impossible to establish benchmarks for the ethics of an AI system.

In addition, it is not straightforward to integrate different types of measure. For example, trustworthy AI requires both a trustworthy system ("trust is justified") and also user trust ("trust is given"). Each of these is a combination of multiple factors. The more AI is implemented, the more complex this will become.

It is not enough to understand just the level of adherence to ethical principles. It is important also to understand the potential impact on business outcomes. AI has already been applied to thousands of use cases and potentially can play a part in any process. So, the number of uses of AI is very large and in each case there may be a different type of impact.

In some cases, this level of complexity may need modelling (eg by using digital twins) to help people to understand the impact of changes.

People and measurement

People have a complicated relationship with measurement. They tend to focus on the measures themselves not the overall intended outcome: “Your performance management system is full of metrics that are flawed proxies for what you care about.” Because of the incompleteness challenge identified above, this is especially relevant to AI ethics. People are also likely to try and 'game' the measurement by distorting or changing the measurement process to show improvements in their performance to the detriment of business outcomes.

Unexpected outcomes

Many ethical difficulties highlighted to date were unexpected, ranging from reputational damage to operational difficulties. As AI supports increasing automation of business processes, there is a risk that AI will both contribute to, and accelerate, the type of event described by Taleb in “Black Swan”. These are events that generate very unfavourable and damaging outcomes (for example, the financial crash of 2007/8) but which are rare (and therefore not present in data).

Therefore, the measurement approach needs to be flexible enough to incorporate unexpected outcomes and their potential relationship with AI implementation. This applies at the technical level but also in terms of detecting anomalies at a higher level. There is a clear role for scenario planning to help leaders to recognise ethical risks early enough to mitigate them.

Measurement complexity

Finally, the difficulty of making the measurements will play a role. Implementing AI across an organisation will necessarily be iterative requiring an iterative (or continuous) approach to measurement. But easy measurement mechanisms may not be available making it difficult to make measurements routine. This will increase the likelihood that measurement will not be frequent enough and that ethical risks will not be identified in time.

A further difficulty is that AI will increasingly be used as a tool to suggest process improvements as part of increasing automation, or as a tool to make measurements. AI will be marking the homework of other AI(s) so at what stage will the complexity become too great for people to understand?

CONCLUSION

Measuring AI ethics issues, their impacts and causes is essential. But measurement is subject to of the following challenges:

- Understanding AI ethics requires human judgement because measures will only provide an incomplete picture;
- Human judgement is not consistent. and cannot guarantee objectivity;
- Root cause of ethical issues may be hard to measure but measuring only what is easy will cause ethical risks;
- Automation using AI may accelerate rare and damaging ("Black Swan") outcomes that any measurement process may not be reliable enough to identify.

In these circumstances, any implementation of measures for AI ethics is necessarily iterative. But alongside a rigorous incremental implementation of the measurement described above, for the foreseeable future there is a clear role for human judgement.

This article was written by Paul Walton, Jasmin Booth, Jamie Rich (Capgemini), Dino Mariutti and Tom Weston (UCL MBA students).

APPENDIX

This Appendix contains more detail about the measures outlined in Figure 2 above. All of the measures need to be business-specific but they need to include the following scope:

- AI implementation;
- AI ethics enablers;
- Business outcomes;
- Feedback and improvement.

There is a table for each of these below providing examples.

AI implementation

Dimension	Requirement	Measures
AI development	• Establish whether the AI is trustworthy and meets the ethical principles (in itself)	• Existence of AI ethics policy for AI development • Level of adherence to the policy and guidelines for "ethical by design" • Maturity of data science measures
AI product procurement	• Establish whether the AI is trustworthy and meets the ethical principles (in itself)	• Existence of AI ethics policy for procurement and commercial processes, applicable standards and accreditation • Level of adherence to the policy
Citizen development	• Establish whether the AI is trustworthy and meets the ethical principles (in itself)	• Maturity of organisational policies and controls on citizen development
AI product management	• Establish whether the AI is trustworthy and meets the ethical principles (in itself)	• Existence of AI ethics policy for product management and other required digital and data skills • Level of adherence to the policy • Degree of implementation of "ethical by design" through each product lifecycle • Existence of, and level of adherence to, standards • Maturity of management of risks • Maturity of testing (against ethical principles and for robustness)
Organisational change	• Establish whether the AI is trustworthy and meets the ethical principles (in itself)	• Existence of AI ethics policy for organisational change • Level of adherence to the policy • Stakeholder surveys (for trust and explainability)

AI ethics enablers

Dimension	Requirement	Measures
Leadership	• Establish whether the leadership is committed, knowledgeably, to the implementation of ethical AI	• Level of leadership knowledge and commitment
Process / service	• Establish whether processes are designed to conform with ethical principles	• Existence of AI ethics policy for process and service design • Level of adherence to the policy
Culture	• Establish whether the organisational culture enables or inhibits ethical AI	• Level of cultural fit with the implementation of AI • The degree to which key values are embedded
Skills & communities	• Establish whether required roles are in place (AI engineering, AI ethicists) • Establish whether other digital and data skills and communities support the implementation of AI • Establish whether the level of data and AI ethics skills and awareness across the organisation is sufficient	• Maturity and sustainability of communities with respect to the incorporation of AI in digital delivery • Level of awareness across the organisation
AI engineering	• Establishing the maturity of • data and AI platform industrialisation • data governance and trust	• Level of maturity
Technology architecture	• Establish whether the technology architecture enables the integration of AI services	• Degree to which the delivery of AI services are included in the technology architecture and governance

Business outcomes

Dimension	Requirement	Measures
Financial	• Understand the actual and potential impact of AI ethics on financial measures	• Maturity of the analytics (see below)
Risk	• Establish whether risks are understood and managed • Establish whether there are mechanisms to provide early detection of previously unforeseen consequence	• Maturity of the risk management process (with respect to AI ethics) • Existence of regular scenario planning
ESG	• Understand the actual and potential impact of AI ethics on ESG outcomes	• Maturity of the analytics (see below)

Feedback and improvement

Dimension	Requirement	Measures
Measurement mechanisms	• Establish whether the quality and robustness of the measurement infrastructure, processes and governance	• Maturity of the measurement mechanisms
Analytics	• Establish the level of capability of the analysis of the measures	• Maturity of the analytics process
AI governance	• Establish the maturity of the AI governance with respect to ethics	• Maturity of the AI governance process
Transformation processes	• Establish whether change requirements are converted into changes reliably	• Maturity of the transformation processes and governance
Variability	• Ensure a sufficiently consistent approach to human judgement	• Noise audits (see Kahneman et al "Noise: A Flow in Human Judgment")

ABOUT US

CAPGEMINI

Capgemini is a global leader in consulting, digital transformation, technology and engineering services. The Group is at the forefront of innovation to address the entire breadth of clients' opportunities in the evolving world of cloud, digital and platforms.

THE UCL SCHOOL OF MANAGEMENT

The **UCL School of Management** is the business school of University College London, one of the world's leading universities, consistently ranked in the global top 20 for its academic excellence and research. The School offers innovative undergraduate, postgraduate, PhD and executive programmes in Management, Entrepreneurship, Business Analytics, Business Information Systems, and Finance, designed to prepare students for leadership roles in the next generation of innovation-intensive organisations.

THE ANALYTICS LAB

The Analytics Lab is an enrichment module for business students where they are able to explore topical questions in the domain of business analytics and digital economy via hands-on experience. Students are offered the opportunity to conduct research and work on projects with leading technology service and consulting companies.

It aspires to help UCL business students and alumni to be in the heart of fundamental changes and digital transformations in the business environment. Students enhance their practical abilities to manage and operate business activities effectively in view of rapidly developing digital and technological advancements in data analytics.