



THE
ETHICS
INSTITUTE



Guidebook to managing The Ethics of AI in Organisations

Kris Dobie & Schalk Engelbrecht

Authors: Kris Dobie and Dr Schalk Engelbrecht

Editorial support: Dr Liezl Groenewald

Cover design and layout: Elsie Weich

Guidebook to Managing the Ethics of AI in Organisations

© The Ethics Institute (TEI) 2025

ISBN: 978-1-0370-7849-1

Published by: The Ethics Institute

Pretoria, South Africa

Website: www.tei.org.za

Contact: info@tei.org.za

© This publication is licensed under a Creative Commons Attribution
| Non-commercial | No Derivatives 4.0 International License.

Disclaimer - This handbook does not constitute legal advice.

For legal advice on compliance with the Companies Act and Regulations or any other legal or regulatory requirements, please consult an appropriately qualified legal advisor.

Transparency statement:

AI was used for research and checking completeness of content. Where AI was used for generating content it is specifically referenced.



**Publications in The Ethics Institute's handbook series
available at www.tei.org.za**

Ethics Risk Handbook (2016) [also available in Portuguese]

Ethics & Compliance Handbook (2017)

The Ethics Office Handbook (2018)

The Social and Ethics Committee Handbook [Second Edition] (2018)

Codes of Ethics Handbook (2020) [also available in Portuguese]

Whistleblowing Management Handbook (2020)

Ethics Ambassador Handbook (2021)

Institutionalising Ethics Handbook (2021) [also available in Portuguese]

Ethics Reporting and Auditing Handbook [Second Edition] (2022)

Ethical Culture Handbook (2022) [also available in Portuguese]

Ethical Leadership Handbook (2023) [also available in Portuguese]

Conflict of Interest Handbook (2024)



THE
ETHICS
INSTITUTE



Guidebook to managing
**The Ethics of AI
in Organisations**

Kris Dobie & Schalk Engelbrecht

Published by The Ethics Institute 2025



Contents

Preface

A. Introduction	1
1. Defining the concepts	3
1.1. What is ethics?	3
1.2. What is AI? Important distinctions	4
1.3. What is AI Ethics?	5
2. Setting the scene	6
2.1. Broader societal issues	6
a) Concerns	7
b) Opportunities	9
2.2. The interface of ethics and AI – Human-centric AI	10
B. The Ethics of AI in Organisations	11
1. Types of AI use in organisations	11
2. AI Ethics risks for organisations	13
2.1. Accuracy	14
2.2. Data security	15
2.3. Bias / fairness	16
2.4. Transparency / explainability	17
2.5. Recourse	18
2.6. Data privacy	19
2.7. Autonomy	21
2.8. Accountability	22
2.9. Job replacement	24



3. Governance and management of AI ethics	25
3.1. Committee oversight	26
3.2. Setting standards (Codify)	27
a) Setting standards for employee use of LLMs	29
b) Setting standards for AI projects: Principles of Responsible AI	30
3.3. Socialise	34
a) Internal Culture Measures	34
b) External Communication	34
3.4. Monitor & Report	35
3.5. AI Project Ethics Risk Management	35
a) Assign responsibility	36
b) Team diversity	37
c) Assess against standards	37
d) Human-in-the-process	41
e) Recourse	42
f) Review regularly	43
C. EU AI Act - Examples	44
D. Conclusion	50
Bibliography	52
About the Authors	56
About The Ethics Institute	57



Preface

We find ourselves at a crucial juncture in history. The decisions we make today - particularly in relation to the adoption and integration of artificial intelligence (AI) - are laying the groundwork for the future in ways that are more immediate and consequential than ever before. These outcomes are not remote; rather, they are concrete, measurable, and unfolding at a speed that necessitates both urgency and responsibility.

The expanding influence of AI within organisations and across society demands a strong commitment to ethics and trust. AI, with all its potential and complexity, does not operate in a vacuum. It interacts with human values, organisational processes, and social norms - prompting critical questions regarding fairness, accountability, transparency, and broader societal impact. These questions cannot be deferred or overlooked. They call for intentional reflection, collaborative learning, and bold yet careful decision-making.

They keyword here is intentionality. Ethical AI will not be achieved accidentally. It is only by being intentional in our ethical approach that AI will serve as a tool for collective good, and that we will avoid unintended harm. To cultivate and sustain trust, it is imperative that organisations integrate ethical considerations into their AI strategies from the very beginning and throughout the lifecycle of these technologies.

While some larger organisations have already made meaningful progress in managing AI ethics risks, it is evident that many others are still navigating their way forward. This guidebook seeks to contribute to that broader effort by offering practical insights and reflections grounded in real-world experiences.

As part of the development of this guide, the authors have engaged directly with several South African organisations to better understand their respective journeys and approaches to AI ethics. These conversations have been instrumental in shaping the content and direction of this guidebook.



The Ethics Institute wants to express its sincere appreciation to the following individuals and organisations for their valuable contributions: FirstRand (Ponyane Mathabatha), FNB (Dr Mark Nasila), Liberty (Tracey Unser, Rekha Naidoo, Inge Rickhoff, and Ruan Schutte), Nedbank (Driekie Havenga), Thungela Resources (Deon van Heerden, Sanet van Schalkwyk, Carinka van der Walt, Johan Conradie, and Franco Holder), Vodacom (Karen Robinson, Cavell Alexander, and Bambo Ntlabati), and Nenzeni Duma.

We hope that this guidebook will assist practitioners and decision-makers to navigate the complexities of AI ethics. Let us always be guided by a commitment to outcomes that advance human dignity and social wellbeing.

Dr Liezl Groenewald

CEO: TEI



A. Introduction

With the launch of ChatGPT on 30 November 2022 the world's eyes opened to the phenomenon of Artificial Intelligence (AI). While AI has been in use long before that in various guises, it was the human-like interaction with ChatGPT that captured people's attention.

Within a very short space of time the use of AI has become pervasive in almost all organisations. Most people use AI in one form or another, as it also becomes integrated into search engines and other IT products.

Organisations are increasingly promoting the use of AI among staff, fearing that they might be left behind as the technology develops in leaps and bounds, and that not using it could impact their competitiveness.



In this rush, it is however critical that the ethical use of AI receives attention.

Many people are concerned about the short- and longer-term impact of AI on humanity. Some of the leading minds in AI development wrote an [open letter](#) already in March 2023, urging a pause on AI training, and for putting in place safeguards for responsible AI development. While there is widespread discussion about developing human-centred AI, this guidebook is not aimed at AI developers.

We have realised that many organisations are still struggling with how to ensure the ethical use of AI by their employees. They are aware that AI has the potential for significant impact, and are aware of ethics risks, but are still finding their way in addressing these issues. The guidebook is therefore aimed at assisting them.

We start out with an introductory section to define the concepts, distinguish between different types of AI, and set the broader scene by focusing on the societal debate around AI.

We then move on to the two questions that many organisations are grappling with:

- What are the ethical risks associated with the use of AI?; and
- What do we, as an organisation, do about it?

We end off with a short section on the European Union AI Act, which is currently the leading legislation on the topic, and gives practical guidance on the ethical risk classification for various AI uses.

This guidebook is intended as a resource for all organisational role-players who have a responsibility to manage the ethics risks associated with AI, including boards, executive teams, ethics practitioners, risk practitioners and IT development teams.

The debate around these issues is constantly evolving. Our purpose is to give a new reader a structured overview of the issues, and practical guidance on how to manage them.

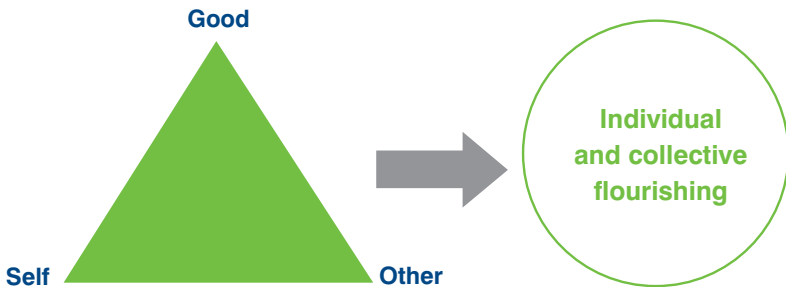


1. Defining the concepts

1.1. What is ethics?

While most readers probably have a good understanding of what ethics means, it is nonetheless worth revisiting this concept before we apply it to AI.

We will use a very simple definition for ethics, which you will find in many of The Ethics Institute's (TEI's) [guidebooks](#). Being ethical means considering not only what is good for oneself, but also what is good for others.



For this guidebook, where we discuss using AI, we however add an additional reflection to the definition: we balance our interests and the interests of others *so that we can achieve individual and collective flourishing*.

So, when we assess whether we are using AI ethically, we should ask ourselves whether it leads to individual (personal) and collective (societal) flourishing. Are we creating lives and societies that could be called “happy”, “good”, or “thriving” – not for machines, but for people on this unique planet?

Ethics is pursued by individuals, but also by organisations. Just as we consider what is good, right and virtuous as individuals, so too do organisations ask how they can balance their interests with the interests of those they have an impact on, and that have an impact on them – from employees to clients, investors, the environment, and the public at large. This is called organisational ethics.



1.2. What is AI? Important distinctions

When most users think of AI, they probably think of ChatGPT or a similar product which can be used to generate writing, code, images or music.

AI is however a much broader field. For our purposes we will limit ourselves to a few concepts.

Artificial Intelligence (AI) refers to the ability of machines to perform tasks that typically require human intelligence, such as learning, reasoning, problem-solving, and perception. It involves using algorithms, data, and computational power to simulate human intelligence, enabling machines to learn from experience and adapt to new situations⁴.

Machine Learning (ML) is a subset of AI. It is a form of “automated learning” where a system improves the accuracy of its own prediction as more data is fed into the system. It can achieve this learning without explicit instruction from humans. Machine learning is therefore useful to see patterns in data that humans were not able to.

Generative AI is again a subset of ML. This is where an AI system automatically generates content like texts or images.

Large Language Models (LLMs) are a subset of Generative AI. These are a form of AI that specialises in the generation of language or text. LLMs are trained on vast amounts of text and data, and are designed to generate coherent language responses. Current examples of LLMs include ChatGPT, BERT, Copilot and Llama.



While there are much more nuanced breakdowns of what AI is, the above is sufficient for our purposes. We will predominantly be referring to LLMs, and machine learning (which is frequently used by organisations for analysing data).

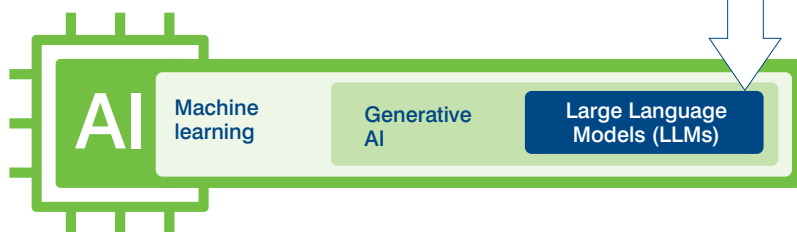
⁴ This is a very simple, but practical definition generated by AI on google search.



The following table summarises these concepts:

	Machine Learning (ML)	Generative AI
Definition	A field of AI focused on analyzing data and learning patterns	A subset of ML focused on generating new data similar to its training data
Main Purpose	Prediction, classification, clustering, decision-making	Creation of new content: text, images, audio, etc.
Is it Creative?	Not usually—focused on analysis	Yes—focused on producing novel or human-like content
Examples	Fraud detection, recommendation systems, speech recognition	ChatGPT, DALL·E, deepfake generators, AI music composers
Training Data Use	Learns patterns for analysis or prediction	Learns patterns to generate similar but new data
Typical Output	Labels, probabilities, decisions, clusters	Images, videos, music, code, or text

Table generated with ChatGPT



1.3 What is AI Ethics?

As you can see from the above explanation of AI, there is a significant amount of data involved when we use AI. For this reason, AI Ethics and Data Ethics are almost inseparable.

Data Ethics refers to the standards, principles and frameworks that ensure the ethical collection, analysis, storage and use of data. Some of the key topics in Data Ethics are *privacy, security, bias, fairness, and accountability*.

AI however has some additional characteristics that have their own ethical implications. Firstly, due to the newfound scale at which we can now engage with data, we can achieve



things that weren't possible before. For example, we can now do facial recognition scanning of an entire street of people, which raises special privacy concerns. But one specific AI feature adds new ethical questions, and this is the automation of decision-making. We can now develop AI which will make decisions affecting people – for example self-driving cars. This raises new questions about accountability.

These specific challenges and risks that arise require us to apply ethical reasoning and principles to these emerging challenges. Attending to these challenges is what is called AI Ethics or Responsible AI, which the International Organization for Standardization (“ISO”) describes as “the practice of developing and using AI systems in a way that benefits society while minimizing the risk of negative consequences”⁵.

We can see that the concept of individual and collective flourishing is again implied in this definition.

Organisations especially need to consider and institutionalise AI Ethics. As organisations prioritise and speed up the embedding of AI systems into internal operations and client services, a lack of ethical sensitivity can lead to significant harms to individuals, organisations and society. It can also break down trust in the use of AI, or in the organisations themselves.

2. Setting the scene

2.1 Broader societal issues

The increasing pervasiveness of AI in society is undeniable. With this comes many social questions that may have an impact on how we live our lives. There are those who are optimistic about such changes, and those who view the changes as potentially dangerous to our way of life.

The overall impact is likely not cast in stone, and the positive or negative impacts will unfold as AI becomes both more powerful, and more integrated in our lives.

It should however be emphasised that this guidebook does not aim to answer broader questions about this impact of AI on humanity. These questions are likely outside of the

² <https://www.iso.org/artificial-intelligence/responsible-ai-ethics>



sphere of influence for most readers. Instead, the intention is for this guidebook to be a practical resource to organisations who are finding their way in the ethical use of AI.

There is however likely a link between what we do at the organisational level, and the overall impact on society. We therefore think some of the broader societal issues are worth a mention – both the concerns, and the opportunities.

a) Concerns

i. Digital divide

There is already a concern that those with access to a good education and to technology have more access to opportunities and resources than those who do not. This divide might expand significantly as technological powers increase. Historian, Yuval Noah Harari, pointed out that the possibility exists that humanity might evolve in two separate species. Those who evolve with AI and technology, and those who don't.

ii. Job losses

Linked to the digital divide is the concern about job losses – also referred to as technological unemployment. More and more companies are replacing workers with AI. It is argued that AI is more cost-efficient and more accurate. For now, the focus seems to be on repetitive tasks, and proponents argue that AI is removing menial rather than meaningful work. Many also state that AI is in fact creating many jobs as well, and that upskilling is key to continued employment. It might however be that those who are less educated or less capable might see jobs and income disappear, while the more educated receive more opportunities.

iii. Manipulated news / facts

A more immediate concern that we already see developing is the 'loss of truth'. Human beings are globally linked through social media and the internet. There are many incidents of deepfake AI images that are surfacing with the intent of changing global sentiment. In many cases it takes an expert eye to determine if something is real or fake. As technology progresses, fake 'facts' might be incorporated in real news or real science, making it difficult to distinguish the truth from fiction.



iv. Societal polarisation

While not inherent to the use of AI, we have over the past decade or more, seen the impact of AI algorithms in polarising society. This generally occurs when algorithms curate the news, advertisements, and social media content that is recommended to individual users. These so-called 'digital echo-chambers' tend to amplify people's pre-held beliefs, thereby widening ideological differences in society. In [certain instances](#) it is even said that social media companies have knowingly contributed to fuelling ethnic conflicts and wars.

v. Environment

The processing power of AI requires a significant amount of energy. The [World Economic Forum](#) estimates that data centres contribute between 3 and 5 percent of US, European Union (EU) and Chinese electricity use, and that demand is currently growing. AI proponents say that AI will help us to find solutions to energy challenges in the longer term, but that requires a short-term rise in electricity use. Time will tell whether the investment pays off.

vi. Accountability

As systems become more complex it becomes increasingly difficult to assign responsibility for the impact of decisions. With AI being more involved in decision-making, it can be difficult to know who to hold accountable for certain societal outcomes. For example, if a doctor is negligent, it is quite easy to hold them accountable. If the same medical mistake is made by AI, who takes responsibility – the manufacturer of the AI, the programmer, the hospital, or the Dr in attendance? The same can be asked for a driverless car that injures a pedestrian - is it the car manufacturer, the AI provider, or the government that did not regulate it?

vii. Existential threat

At the extreme end of concerns is that AI might become so powerful that it, rather than humans, takes over control of humanity. Very much in the same way that humans have out-evolved other species, AI might out-evolve humans. We will then be at the mercy of whatever type of AI we have created. Data scientists are at odds about whether this is a serious concern, and if it is, what the timelines and potential impact might be. The views differ vastly. Dependent on who you listen to:

- It is a risk to the future existence of humanity – https://en.wikipedia.org/wiki/Existential_risk_from_artificial_intelligence
- It is a risk to what it means to be human – <https://www.scientificamerican.com/article/ai-is-an-existential-threat-just-not-the-way-you-think/>
- It isn't a risk at all – <https://medium.com/data-science/existential-risk-from-ai-a-skeptical-perspective-35f0cd7c9fa4>

b) Opportunities

There are, however, also many AI optimists who believe that with improved computing power comes improved problem-solving abilities. In this way, AI might assist humanity with solving some persistent societal issues.

As AI develops and quantum computing gives us more powerful tools, scientific progress will likely increase exponentially, especially in areas like improved healthcare. AI can help us find medical solutions that have evaded us for centuries.

AI is also seen as a potential game changer in education, where it can create personalised learning to suit each individual learner. In this way it may give education opportunities to many who did not previously have them, and could assist in reducing the digital divide.

For most of the concerns about AI listed above, there are also those who believe that the impact of AI will be positive, rather than negative in these areas. For example, while some fear misinformation, others believe that AI might help us to better identify fake news, and thereby deepen democracy. Some fear environmental degradation, while others believe AI will help us overcome this innate challenge to humanity. Job losses may be a concern for some, but others argue AI will give us more meaningful jobs in the long run, much like what happened with the industrial revolution.

The [European Parliament](#) argues that one needs to find a good balance between strict regulation and accountability on the one hand, and allowing enough freedom for innovation on the other. Inasmuch as they see the potential risks of AI, they also see the underuse of AI as a potential threat to the EU. If the EU falls behind in harnessing AI opportunities, they might become less competitive internationally.

The same logic would also apply to businesses and other organisations. While we should avoid ethical pitfalls, we should be careful not to fall behind the curve of technology.



As you can see from the above discussion – the jury is still out about whether AI will have an overall positive or negative impact on humanity. For this reason, we should keep focusing on human-centred AI to ensure that the positives are realised, and the negatives are avoided.

2.2. The interface of ethics and AI – Human-centred AI

Any use of AI is based on data, and large volumes of it. With the excitement that the innovative use of AI brings, there is the risk that the people behind these volumes of data fade from view. The data at our disposal today – from transaction records to GPS logs, contacts, online search histories, social posts, heart rates and sleep rhythms – represents seemingly limitless possibilities for learning, trend recognition, prediction, and even behavioural priming. But behind every piece of data is a person with preferences, desires, feelings, vulnerabilities, dignities and rights. Here enters ethics, and the need for human-centred AI.

Human-centred AI is a commitment to prioritise people and their wellbeing in the development and use of AI. It aims to leverage the opportunities that AI offers to foster human dignity and promote flourishing, while avoiding or minimising the associated risks. Using AI to prompt unwitting citizens to vote in a particular way in a national election, for instance, undermines autonomy, fails to demonstrate respect for people, and erodes our democratic values. Using AI for the early detection and treatment of diseases like cancer, on the other hand, supports human dignity, improves the wellbeing of citizens, and strengthens the character of a society.

The ethical and human-centred use of AI requires more than compliance with law. Adherence to the law is the least an organisation can do to use AI responsibly, but it is woefully inadequate if our aim is human-centred AI, especially if one considers how far the law lags behind technological developments.

Achieving human-centred AI must therefore be a deliberate commitment to use AI in specific ways, even when not required by law. It is, however, also a balancing act. A human-centred approach means using AI in ways that help individuals to realise their self-chosen goals, while not making people redundant; that enlarge our capabilities, without removing our responsibility; and that lead to cohesion and collaboration between people, rather than polarisation and division⁶.

6 Cf Floridi et al. 2018. [AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations](#). *Minds and Machines* 28, pp. 689 – 707.



B. The Ethics of AI in Organisations

1. Types of AI use in organisations

AI has been around for some time, but it is especially after the launch of ChatGPT in 2022 that it has become intertwined with our daily lives. No longer are we merely subject to the workings of AI in our personal lives (for instance, through our personalised social media feeds, or the music recommendations we receive based on our listening history); but today ordinary employees also progressively use AI tools in the workplace for everyday enquiries posed to so-called “digital assistants”, for automated minute-taking during online meetings and summarising and drafting reports.



Organisations have however been using AI for specific 'project' purposes since before the proliferation of LLMs, and this use has simply been upscaled with improved AI abilities.

Organisational project use of AI includes using it externally for customer service (for example, the use of chatbots to answer customer queries), and internally for things like fraud detection in financial services and insurance, credit rating for loans, and screening candidate CVs in the appointment process.

For the purposes of this guidebook, we find it useful to distinguish between two kinds of AI use in organisations:

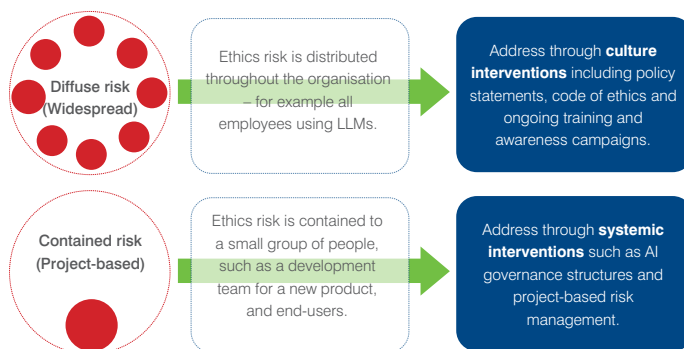
1. The widespread use of AI tools by most employees – integrated in their everyday work (mostly using LLMs);
2. The project-based use of AI tools to optimise the organisation's processes (either using LLMs, or ML).

Both of these uses may have ethical implications. An employee who uses ChatGPT to draft a report may be breaching confidentiality or privacy rules; an algorithm used in recruitment to screen suitable candidates might be biased; and an AI tool used to determine the creditworthiness of loan applicants could inform decisions unfairly.

Distinction between the widespread employee use, and specific project-based use of AI, is quite important for two reasons. Firstly, many organisations are not yet that technologically advanced that they will have project-based AI uses. They are more likely concerned about how their staff use LLMs such as Copilot or ChatGPT.

Secondly, the way organisations would address these risks is quite different. The widespread use of LLMs would require more culture-based interventions, because the risk is spread out across the entire organisation. Project-based use of AI would require systemic risk governance interventions, with formal project reviews, as well as culture interventions for those employees who use the specific AI tool.

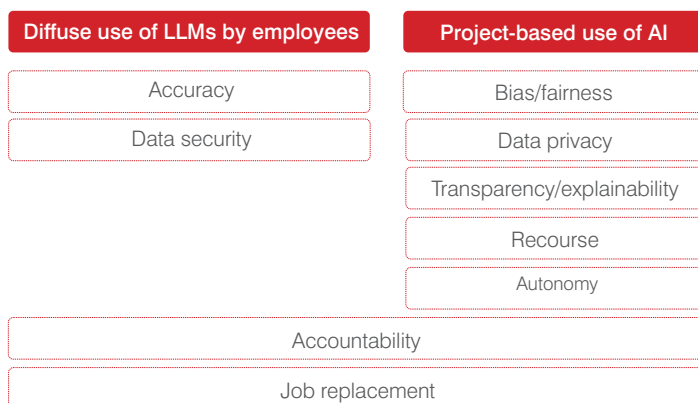
The following graphic sets out this distinction:



2. AI Ethics risks for organisations

As AI is being rolled out in organisations, there is an increasing number of examples of when things went wrong. We believe there is much to learn about ethics risk from past experience and will continually refer to examples to promote understanding of otherwise abstract issues. It should be noted that most of the companies mentioned did not necessarily act in bad faith, but we share their stories so that it can assist other organisations to avoid similar errors.

The following is a discussion not only of obvious errors, but also of ethical issues that organisations will need to consider as they increasingly use AI. It is again possible to draw a distinction between the diffuse use of LLMs by employees and the project-based use of AI. There will be overlaps between the distinctions, but we nonetheless believe there is practical value at this stage to separate the issues.





Let us start with the main issues applicable to LLMs:

2.1. Accuracy

What is the issue?

There are a number of cases of lawyers using AI to do legal research, and AI ‘hallucinating’ false case law. In other words, the AI simply made-up case law to support the lawyers’ cases. One court ruled that lawyers must always highlight when they use AI so that case law can be checked, and another ruled that the lawyers remain professionally accountable for the case law they present. In one court lawyers were fined for presenting false case law.

These lawyers were simply ‘caught out’ because there were opponents checking their work. It is almost inevitable that mistakes happen in all workplaces.

The fact is that LLMs are not yet 100% accurate. This is because an LLM is not linked to a database of facts which it searches for your answer. So, it is not quite like using an internet search engine. LLMs work by predicting the next logical word (or token) based on the words that came before. They are therefore more like very sophisticated text predictors than search engines. They are however trained on vast amounts of data, so the likelihood that they will give you accurate information is quite high. However, the exact information that you are looking for might not have formed part of its training, leading to inaccurate information.

The issue of LLM inaccuracy is so prevalent that it has been given a name: ‘hallucinations’. In some situations LLMs will simply make up facts. If it were human, it would probably be deemed to be lying, but since it has no moral agency, it is simply said to be hallucinating. A 2024 study on using LLMs in academic writing found that ChatGPT 4.0 hallucinates as much as 28% of references.

The problem is that humans have been trained to believe the answers that we are given by computers through search engines such as Google or Bing. The other problem is that LLMs are accurate almost 90% of the time, so we are caught off guard when they make things up. They also don’t indicate when they don’t know – they simply give us inaccurate information with the same confident tone as accurate information.



How can organisations address this?

The first step is policy clarity on who remains responsible for the output of work, placing the onus on employees to be professionally responsible for the work that leaves their desks.

This should be followed by training and awareness campaigns, clarifying the risks, guidance on how to reduce the risks, and what the organisation's policy is.

For project work, organisations should be careful of releasing AI products that have not been tested. Also keep in mind that many issues might only appear after release, so continuous monitoring and testing is critical.

2.2. Data security

What is the issue?

One of the greatest risks that organisations face is that their staff use 'open' LLMs which are not ringfenced to protect the company's data. This means that if your employees enter confidential client information, or proprietary company information onto the system, this becomes part of the LLM's training, and the confidentiality is lost.

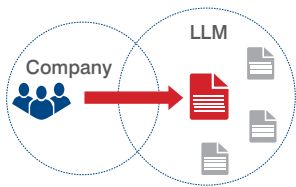
It may not be possible for others to access the entire folder with your information, but your data integrity is breached nonetheless. This can be especially problematic if your organisation has source code that is part of its intellectual property. If the LLM is trained on this code it can become common knowledge.

The same would apply to confidential client information, patient information, or HR information. If prompted in the right way, your personal information may already be on some LLMs.

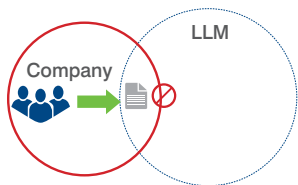
How can organisations address this?

The first step is to raise awareness with employees about the risks of using unprotected LLMs. This can be done through policy statements, training and awareness campaigns.

Secondly, the company should provide staff with safe options to use. This can be done through subscription or enterprise versions where the company opts out of using its data for LLM training.



This graphic illustrates how confidential company data can become part of the training material for LLMs if staff upload it onto standard open LLMs.



With subscription, or enterprise versions of LLMs it is, however, possible for companies to opt out of their data being used for LLM training. This has to be specifically agreed with the provider. In this case the data is ringfenced and protected.

Some however argue that these protections are not failproof, and if one has particularly sensitive information it should not be put onto LLMs.

Data security for large databases of information is a security concern, even when not uploaded onto LLMs. How this is dealt with would need technical knowledge of data specialists and is beyond the scope of this guide.

Let us now turn to issues that are more prevalent in project-based AI work.

2.3. Bias / fairness

What is the issue?

Let's say you develop a system that guides who gets into a university and who does not. Traditionally, pupils from private schools do better, and more of them are placed in universities. The algorithm is trained to look for patterns to make smarter recommendations. It then picks up on this link, and therefore places pupils from private schools before placing students from state schools. Because it was trained on biased information, it ends up making biased recommendations.

This scenario is based on a real case from the UK. Now you can imagine, if your child has done just as well as a private school candidate, but is pushed to the back of the queue, you would feel they were treated very unfairly.



Many other examples of biases exist:

- An Amazon recruiting algorithm was trained on more male than female applications and ended up favouring men.
- The US court system uses an algorithm to predict whether a defendant is likely to re-offend. [Research](#) has shown that, because of the data it was trained on, the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm predicted twice as many false positives for black offenders as white.
- When banks do credit assessments to determine whether people qualify for loans, or their interest rates, there seems to frequently be bias against black applicants – most likely due to biases in the training data.

As you can see from the examples, we use AI for very important decisions. We use it to decide how people are sentenced in courts, who gets jobs, and who is given access to loans. These decisions can have a significant impact on people's lives and livelihoods, and if there is unfairness built into the system it will prolong and exacerbate societal injustices. Such outcomes will also break down trust in AI, and in the organisations using them.

How can organisations address this?

There are various reasons why bias might appear in an AI system. The most common is that we use historically biased data to train the AI. These biases are then perpetuated by the system.

Machine learning seems especially prone to make biased decisions, and developers need to always ensure that they keep this in mind and design to remove bias, and keep testing to ensure there is no bias. One organisation interviewed has a rigorous pre-launch testing, and then re-tests each system every two years.

The solutions to remove bias are generally more technical. It requires risk management processes and the resources to specifically test models for bias. There are many [resources](#) on the internet to give guidance in addressing this issue. Practitioners that we spoke to in our research also indicated that using diverse development teams can contribute to less bias in products.



2.4. Transparency / explainability

What is the issue?

Imagine you apply for a home loan, and you are turned down. When asked why, you are told it was done by AI and they don't know how it came to its decision.

This is the so called 'black box' problem of AI, where decisions are made, but cannot be explained because the algorithms that are used are unclear. If you cannot explain how you came to a decision, it is very difficult for individuals to determine whether the decision is fair, or even accurate, and it therefore breaks down trust.

For this reason, many are arguing for a 'glass box' approach, where AI gives clear and transparent reasons for the decisions it makes – leading to a movement towards Explainable AI (XAI).

How can organisations address this?

Explainability and transparency should be a consideration when developing any machine learning AI product. Addressing the risk would require technical solutions. Ensuring transparency of decisions would therefore need to be written into policy, and built into risk management processes for the development of AI products.

2.5. Recourse

What is the issue?

As is indicated in the above section, AI does, and likely will continue making mistakes that affects groups and individuals.

This is not dissimilar to what was happening before AI. Organisations make mistakes.

What is however critical is that there is some means for individuals to have their concerns resolved. As more and more customer service queries are attended to by chatbots, it becomes more and more difficult for customers to have unique problems addressed.

How can organisations address this?

At some point there should still be a human in the system that takes final accountability and who can be contacted to address issues that are not otherwise solvable.



2.6. Data privacy

What is the issue?

Machine learning is generally applied to large datasets. This can be clients' personal and sensitive information, patient information, staff/HR information, or company proprietary information.

There is therefore an inextricable link between data ethics and AI ethics. Many questions of AI ethics relate to how organisations protect, save and utilise the vast amounts of data that they keep.

The following are potential pitfalls:

- Gathering data through unethical means;
- Buying third party data that was collected unethically or illegally;
- Using data for purposes other than what it was collected for;
- Not keeping data safe; and
- Selling data without consent.

While these pitfalls seem easy enough to avoid, it can be quite tempting to go against these principles. When companies hold large quantities of personal information, it can help them in more personalised marketing or product design.

Think of the following scenarios:

- A medical aid holds data of all your medical treatments and medication.
 - o If they are also in life insurance, are they allowed to use this data to determine patients' risk profile and by implication their life insurance premiums and exclusions?
 - o Can they use this data to warn patients about possible health threats?
- A company develops a new app that requires vast amounts of personal information to use it effectively.
 - o Can they force their employees to subscribe to this app and thereby give up their personal information?

In most cases it is about being transparent with your clients and obtaining permission for using information for specific purposes. This should be done transparently and not hidden in the small print. You should also not coerce people into giving up their personal information.



FURTHER READING

The Five Ps of Ethical Data Handling

1. Provenance: Where and how was the data obtained?
2. Purpose: Are you using the data for the purpose for which it was obtained?
3. Protection: Is the data safe from theft or abuse?
4. Privacy: Are we respecting the privacy of individuals?
5. Preparation: Are we ensuring sufficient care in using the data?

Source: Harvard Business Review

<https://hbr.org/2023/07/the-ethics-of-managing-peoples-data>

Example:

A well-known example of overstepping privacy lines is from the US where the retail company, Target, was tracking customer purchases as many retailers do. Based on these purchases it would then recommend other products that individuals might be interested in. Building on this practice, Target started tracking purchases of specific products that would indicate that a customer is pregnant. When triggered, they would start marketing pregnancy and baby products to them. The problem arose when the father of a 17-year old girl found the marketing material addressed to her. Unaware of his daughter's pregnancy, he was upset that the company would promote premarital sex to minors. For many, pregnancy is a very private issue, and people may get upset when this privacy is not respected.

How can organisations address this?

Personal data is quite well-regulated. South Africa has the Protection of Personal Information Act (2013), and the EU has the far-reaching General Data Protection Regulation (GDPR) of 2018. The principles of this legislation should be carefully considered in all aspects of data management.

As the example however illustrates, one needs to be sensitive to how people may see the infringement on their privacy, even when the use might be legal.



2.7. Autonomy

What is the issue?

The term 'nudging' is used for influencing people's decisions and behaviours. Health insurance companies are, for example, nudging their customers to remain healthier by giving them rewards for exercising. This is a form of positive nudging. Nudging can however also have a negative impact on the consumer – for example, placing sweets, snacks or other unhealthy options in the check-out aisle where people spend more time.

In apps, nudging is often achieved by highlighting certain options (for example making it easy to purchase something), and hiding other options (like making it difficult to end a subscription). One can see how this can be perceived to be manipulative and impact on the free will of the person who wishes to make their choice.

Big data opens endless opportunities for nudging, but also poses a number of ethical questions.

One of the biggest data scandals to date is the Cambridge Analytica incident of the 2010's. In short, Facebook allowed the company Cambridge Analytica to gather personal data of Facebook users and their connections. This data was used to profile people to get a sense of their political views. Based on this, different information was sent to different voters to influence their behaviours. It was used by Ted Cruz and Donald Trump in their 2016 presidential election bids. This was predominantly a scandal of data privacy, but also asks questions about how masses of data can be used to influence people's behaviour. As we indicate in section A.2.1.a) on page 8 above, data can be used to polarise society.

The EU AI Act, which is further discussed in section C, only specifically prohibits eight uses of AI. One of these is the subliminal, manipulative, or deceptive use of AI, such as misrepresenting facts to vulnerable people to lure them into gambling.

Article 5 - Prohibited AI practices

1. *The following AI practices shall be prohibited:*

(a) the placing on the market, the putting into service or the use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm.

Extract from Recital 29: *Such AI systems deploy... manipulative or deceptive techniques that subvert or impair a person's autonomy, decision-making or free choice in ways that people are not consciously aware of those techniques or, where they are aware of them, can still be deceived or are not able to control or resist them.*

How can organisations address this?

One could argue that the law here provides a minimal guideline, and a higher ethical standard of transparency and non-manipulation should be applied in practice.


The first thing is for organisations to define their values and be clear about their stance on manipulative content. Secondly, these standards then need to be built into AI project governance as will be further explained below in section 3.

The following issues might be relevant to both diffuse use of LLMs, and project-based AI:

2.8 Accountability

Example:

In a Canadian case, a passenger asked Air Canada if he could get a discount for travel to his grandfather's funeral. Air Canada has a policy on 'bereavement travel' allowing this. During his booking he spoke to a chatbot on the Air Canada website. The chatbot said he could submit his claim for a refund within 90 days after the trip. When he tried to do so, he was however told that he was supposed to apply before purchasing the ticket – as is set out on another part of their website. He had taken a



screenshot of the advice from the chatbot, and when they wouldn't reimburse him, he took them to court. Air Canada argued that the chatbot was a separate legal entity and that they were therefore not responsible for its actions. The court disagreed, saying they are responsible for any information on their website. They were ordered to pay the difference in airfare and other costs.

What is the issue?

A very important question for organisations to understand, is “Who is accountable for AI’s decisions?”

This plays out in two spheres. The first is within the organisation. If an employee uses AI that produces inaccurate information, is the AI accountable, or the employee?

The second is the public sphere – in other words, the organisation in society. If the company’s AI gives the wrong advice, or causes harm, who is now accountable – the company, the employee, or the AI?

In short, it is possible for AI to make mistakes. This can include inaccuracies, hallucinations, or biased decision-making. It can even include guidance errors in self-driving vehicles, leading to injury or death. The impact on individuals and groups can be significant.

Issues can also vary in complexity.

In a very simple example, an employee uses a LLM to draft a report but does not check the accuracy of the information obtained. Consequently, a client receives inaccurate legal or policy advice. The emerging consensus seems to be that while employees use AI as a resource, they are still professionally responsible for the information that leaves their desk.

Now imagine that the company rolls out a much larger project, and in time the employee is replaced by AI. This more autonomous AI gives clients advice that is not only unethical, but illegal. The clients act on this advice to their own detriment.

The problem with diffuse accountability is that it is extremely complex to determine where things went wrong, and consequently, who has to be held accountable. Recent court cases seem to indicate that the company that deploys the AI remains accountable for its use.



How can organisations address this?

To address the lack of accountability in using LLMs, many organisations have a policy, or an inclusion in the Code of Ethics that stipulates that people remain accountable for the work that leaves their desk, regardless of how it has been produced.

For larger AI projects, many organisations assign a specific project champion to take overall accountability when a new AI project is launched. Everything from adhering to timelines, project profitability, fitting into institutional architecture, policy adherence, risk management and ethical considerations becomes their responsibility. This ensures that the right team is assembled, and the right questions are asked along the way.

2.9. Job replacement

What is the issue?

Many organisations are already finding that AI and LLMs can do the work previously performed by a large number of employees. Those who do structured routine tasks are more likely to be replaced. This can include check-out staff at retail outlets, call centre agents, and those who do basic data capture and analytics.

At this stage most of the jobs being replaced by AI are considered mundane and repetitive. As AI progresses and quantum computing improves, it is however likely that more and more complex jobs will also be replaced, which will be cause for organisations to ask more difficult questions about where they will draw the line.

One bank indicated that their clients are people, and they are of course quite concerned with keeping their client pool employed – including their own employees. It may however become quite difficult for organisations who do not automate tasks to remain competitive. In this way it may very well be a race to the bottom.

It is at this stage not clear to what extent job replacement will be a concern. Some argue that new types of jobs will be created, which may even be more stimulating. But it is not clear if these jobs will be accessible by those who previously performed more repetitive tasks.

How can organisations address this?

This is a conversation for the board and executive management team to have before they are faced with the immediate decision. There are strategic issues at stake, and the

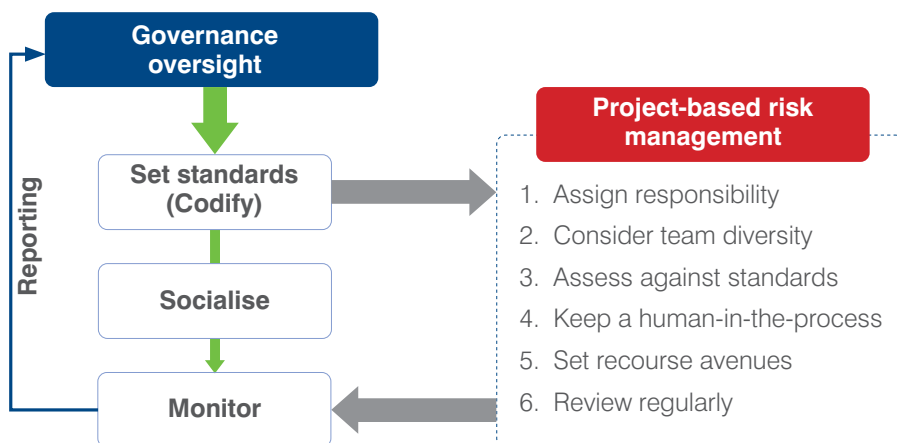


long-term integrated value creation of the company should be kept in mind. The overall purpose of human-centred AI, namely prioritising human needs, values and capabilities in AI development, should be considered.

3. Governance and management of AI ethics

Organisations should follow a proactive and systematic approach to the governance of AI Ethics. This is the responsibility of the Governing Body through its subcommittees. Depending on the sophistication and scope of its AI use, an organisation may decide to direct and oversee AI Ethics through existing committees, e.g., the Risk Committee, or through a dedicated technical subcommittee, e.g., an AI Oversight and Deployment Committee.

Those who oversee AI Ethics must attend to both the widespread use of AI tools by employees as part of their daily work and the project-based development (or purchase) and deployment of AI systems. Effective governance will include general standard-setting (*codification*), organisation-wide training and awareness of these standards (*socialisation*), and ongoing control of AI use in the organisation (*monitoring*), as outlined in the illustration below. Once an organisation starts implementing its own AI systems, oversight will also include project-based ethics risk management (see below).





3.1. Committee oversight

Establishing oversight of AI Ethics aligns an organisation with the principles and recommended practices of the King V Report on Corporate Governance Draft. King V Draft makes specific provision for the governance of emerging technologies like AI and includes recommendations to ensure the trustworthy and ethical use of such technologies (see textbox below⁴).

The committee that oversees AI Ethics should be diverse and have a combination of the ethics, technical, legal, and risk management competence necessary for the provision of effective oversight. The committee is responsible for, among other things:

- Setting the standards that will govern the widespread and project-based use of AI in the organisation;
- Tracking the approved AI tools and acceptable uses in the organisation;
- Ensuring the necessary steps are taken to create and maintain a culture for the ethical use of data and AI in the organisation;
- Providing oversight of project-based AI governance, including the approval of project concepts (based on impact and risk assessments), the go-ahead for the testing and deployment phases in the AI lifecycle, and ongoing monitoring of system performance and the control environment;
- Providing oversight of internal and external audits of AI performance and its associated control environment.

King V Draft and Information Governance

The King V Report on Corporate Governance Draft makes specific provision for the governance of data, information and AI technology. It includes the following principles and recommended practices.

Principle 9 (Information Governance)

The governing body governs information and its deployment through technologies to enable the organisation to expand its opportunities and set and achieve its strategic objectives.

4 To be updated, if necessary, after the publication of King V.

Emerging technologies as sub-set of information governance

With regards to the oversight of emerging technologies, such as artificial intelligence and machine learning, the governing body, or the committee delegated to, should ensure that arrangements are in place to safeguard, among others:

- a) Ensuring that every AI system that is deployed (including bought, built, used or sold) by the organisation adheres to appropriate levels of ethical and trustworthy characteristics.
- b) Ensuring that all processes, resources and tools used to develop, implement, and manage AI systems in the organisation are subject to human and related oversight mechanisms, including:
 - i. The level of oversight or intervention is aligned with the severity of the risk involved for the organisation or a third party.
 - ii. Identification of areas where human intervention is a requisite as well as transparency about AI potentially affecting third parties without human intervention.
 - iii. Ongoing oversight of AI systems which perform continuous learning and change behaviour to ensure that these systems remain to be deployed and used responsibly.

3.2. Setting standards (Codify)

One of the first steps for the effective governance of AI Ethics is standard-setting. An organisation should make its strategy, approach and standards for the development and use of AI clear through codes, policies, and guidelines. These standards will set the tone for AI use, guide the decision-making of employees and technical development teams, provide criteria for distinguishing between appropriate and inappropriate use of AI in the organisation, and communicate the organisation's values and approach to internal and external stakeholders.



Three kinds of documents are useful in this regard:

- (1) An Employee Acceptable Use of AI Guideline
 - This should also be incorporated in the organisation's Code of Ethics
- (2) A Data and AI Ethics Policy, and
- (3) A Risk Management Standard for the Use of AI.

The purpose of each is explained below:

Key Ethics Guidance Documents for the Ethical Use of AI

Employee Acceptable Use Guideline: A short and accessible document explaining the acceptable and unacceptable informal uses of AI tools (particularly the use of LLMs) by employees as part of their everyday work. This would include for instance the prohibition against feeding confidential or proprietary company or client information into public LLMs like ChatGPT, or passing off AI generated text as one's own work. An example is given below in section a).

AI Ethics Policy: A slightly more detailed policy providing guidance to those in the organisation participating in the development and deployment of AI tools (or the procurement and use of third-party AI tools). Such a policy would include the organisation's strategy and approach to AI, the principles of responsible AI Use (Section b) below), roles and responsibilities relative to the development and deployment of AI tools (including those of Business Units, Compliance, Internal Audit, Risk Management, and Oversight Committees), additional resources (for instance, references to relevant legislation and regulation), and key contacts.

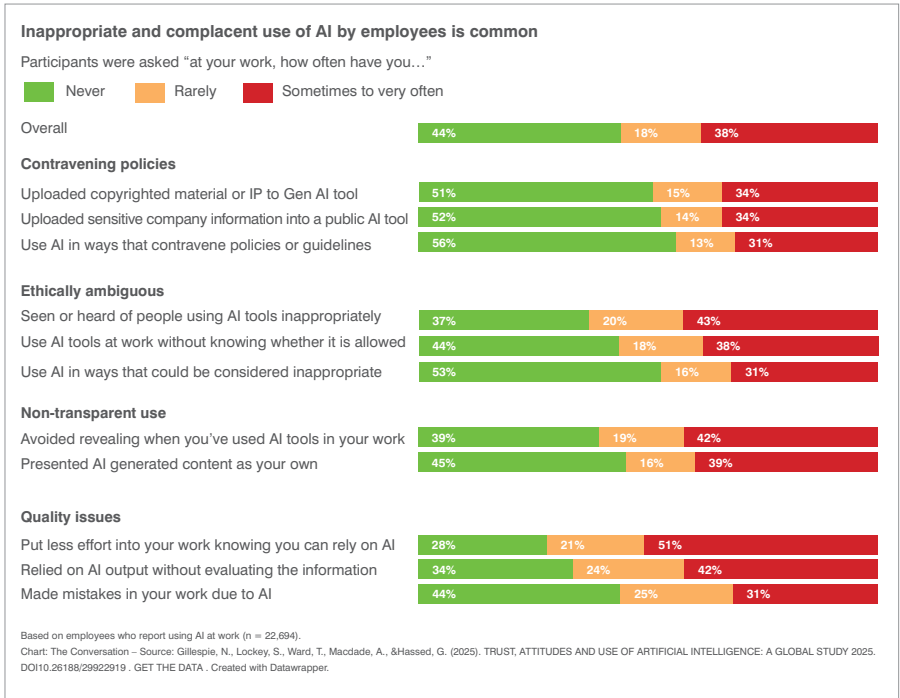
AI Risk Management Standard: A substantial document detailing the possible uses of AI, the Risk Management Approach, and the Risk Methodology. It will therefore include risk management steps required at each stage of the AI lifecycle, from Project Proposal to Data Collection, Model Development, and Deployment and Monitoring. The Standard will outline the required risk-, impact- and bias assessments, an AI risk catalogue, the timing of approvals, ongoing risk and model development documentation, required monitoring, and required reporting.

In the two subsections below, we provide more information on the possible content of the Employee Acceptable Use Guideline (governing widespread employee use of AI) and the AI Ethics Policy (governing the project-based development and deployment of AI systems).

a) Setting standards for employee use of LLMs

All organisations – even those that are not developing or using AI systems – should be wary of the inappropriate use of LLMs by employees. Such use opens organisations to reputational harm, financial loss, and legal liability.

A [recent study](#) conducted in 47 countries and involving over 30 000 employees found that 58% of people regularly and intentionally use AI at work. Almost half of these employees admit to using AI incorrectly⁵. The risky ways employees use AI tools include “uploading sensitive information into public tools, relying on AI answers without checking them, and hiding their use of it”⁶. More examples of inappropriate and complacent use, and their frequency, are listed in the chart below:



5 Gillespie, N., Lockey, S., Ward, T., Macdade, A., & Hased, G. (2025). [Trust, attitudes and use of artificial intelligence: A global study 2025](#). The University of Melbourne and KPMG. DOI 10.26188/28822919. Available at: [Trust, attitudes and use of artificial intelligence](#) [Accessed 1 May 2025].

6 Gillespie, N. & S. Lockey. 2025. “Major survey finds most people use AI regularly at work – but almost half admit to doing so inappropriately” in *The Conversation*, 29 April 2025. Available at: [Major survey finds most people use AI regularly at work – but almost half admit to doing so inappropriately](#) [Accessed 1 May 2025].



Given the frequent misuse of AI by employees, and the study's finding of a "significant gap in the governance of AI tools", it is prudent for any organisation to formulate clear guidelines (Do's and Don'ts) for the appropriate use of AI tools, to communicate these guidelines, and to train employees on them. This can improve the AI literacy of employees, and can help secure the gains in productivity and innovation that AI tools promise, without exposing the organisation and its employees to risk.

Here are some guidelines for the ethical use of AI tools at work:

Ethical Use of AI Tools at Work: Guidelines for Employees (Sample policy)

Artificial intelligence (AI) provides useful tools that present many opportunities for increased efficiency and innovation. It does however require responsible use to ensure unwanted outcomes.

Be aware that AI can make mistakes and you should check your work. Also know that if you upload data onto an AI tool (such as ChatGPT or Copilot) it can divulge personal information or company proprietary data.

When making use of AI, ensure that:

- The information produced is accurate; and
- Data privacy and intellectual property is respected and protected.

Never:

- Dishonestly pass off the work of AI as your own;
- Allow decisions impacting on people to be made by AI; and
- Use AI that was not approved by the IT department.

Always remember that you are accountable for the work you produce using AI.

b) Setting standards for AI projects: Principles of Responsible AI

Standard-setting also includes agreeing on the principles that will govern the development and deployment of AI systems, once an organisation decides to move to this stage of the use of AI. These principles attempt to answer the question "Which core ethical ideas should guide the thinking and decision-making of project sponsors and technical teams throughout the AI lifecycle?".

The AI lifecycle

According to UNESCO (2021) the AI lifecycle ranges “from research, design and development to deployment and use, including maintenance, operation, trade, financing, monitoring and evaluation, validation, end-of-use, disassembly and termination”.

AI Ethics resembles and overlaps with the ethics of many other practices, including journalism and research. To the extent that LLMs like ChatGPT provide people with information that they use for decision-making, AI developers need to ensure, like journalists, that the information their AI tools provide is both accurate and responsible. To the extent that machine learning is a form of research, leading to conclusions and predictions based on surveyed data (often attached to human subjects), the developers of AI models must adhere to research ethics and ensure that the people whose data is being used are informed, consent provided, and are not harmed in the process (see the Facebook Emotional Contagion Case Study)⁷.

Case Study: Facebook and Emotional Contagion

In 2014, a study was published demonstrating so-called “emotional contagion” on social media. Researchers found that when Facebook users were confronted with more negative content in their News Feed, they were more likely to post negative status messages, while more positive News Feeds led to more positive statuses. The conclusion of the study was therefore “that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness”.

While this research is at once interesting, potentially significant, and almost certainly legal, it immediately sparked intense moral controversy. The reason? Facebook and the researchers had not only observed existing social media data, but they had also influenced it. By deliberately modifying the feeds of almost 700 000 Facebook users for a week in 2012, the study had in fact set out to manipulate the emotional states of people, without their awareness.

The study raised questions about *informed consent* and the ethical *sourcing of data*. But even if the study was approved and legal, it also led to an erosion of trust between Facebook and its users.

⁷ See Meyer, R. 2014. “Everything We Know About Facebook’s Secret Mood-Manipulation Experiment” in The Atlantic (24 June 2014). Available at: <https://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secret-mood-manipulation-experiment/373648/>



The set of principles most closely aligned to AI development, however, is found in the field of healthcare- or bioethics. The developers and deployers of AI systems should be guided, like medical doctors and other healthcare practitioners, by the core principles of *beneficence*, *non-maleficence*, *autonomy* and *justice*. In other words, our uses of AI should benefit as many people as possible; should “do no harm”; should respect people’s ability to make their own choices; and should promote equality and avoid any forms of discrimination.

Due to the unique nature of AI, however, it is necessary to add additional guiding principles, including *explainability*, *oversight* and *accountability*.

In the table below, we list and explain some of the most prominent principles to govern the use of AI⁸.

Principle	Description
Beneficence (“Do only good”) ⁹	Developers of AI systems should seek to improve society, serve the public interest, and promote the wellbeing of as many people as possible. Implicit in this principle is the aim to ensure that the benefits of AI technologies are shared by all those affected by it.
Non-maleficence (“Avoid harm”)	The flipside of beneficence is the need to consider and avoid the possible harms associated with AI technologies. Through misuse or overuse, AI can threaten privacy, autonomy, and equality. Non-maleficence therefore includes preventing bias and discrimination, and putting secure constraints on the use of AI.

8 These principles are synthesised from a variety of sources, including: Floridi et al. 2018. [AI4People – An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations](#). Minds and Machines 28, pp. 689 – 707; Institute and Faculty of Actuaries and Royal Statistical Society. 2019. [A Guide for Ethics Data Science](#). Available at: [An Ethical Charter for Data Science WEB FINAL.PDF](#); UK Finance and KPMG LLP 2019. [The Ethical Use of Customer Data in a Digital Economy](#). Available at: [The Ethical Use of Data in a Digital Economy](#).

9 Note that in our list of Principles we have subsumed justice under beneficence and non-maleficence by including the sharing of benefits and the promotion of equality under these principles.



Autonomy	Autonomy is the ability and the right of people to make their own decisions. It is a recognition of the intrinsic value of human choice. To safeguard autonomy, it is important to keep people informed of the collection and use of their data. When AI systems include a component of behavioural priming or “nudging”, this should be apparent to the users or targets of the system. And, because AI often entails the automation of decision-making, promoting autonomy in AI means guarding the ability to choose when decision-making is delegated (promoting human agency), while maintaining control over AI systems (restricting AI autonomy).
Competence and due care	AI developers must apply best practices, use robust algorithmic methods, understand possible sources of error and bias, and regularly review their models.
Transparency	Transparency applies to both the development and deployment of AI systems. It means engaging with diverse stakeholders at the outset of an AI project, and also being open and honest about why, when, and how decisions are delegated, and what the potential risks and benefits are.
Explainability	When decision-making is delegated to AI systems, the conclusions drawn by the system must be explainable and intelligible (one must be able to answer why and how a judgement was made). AI systems are notoriously unclear, sometimes referred to as “black boxes”. Full explainability is therefore not always possible. In such cases developers must still be able to explain the rules used by algorithms, and where explainability becomes impossible, to maintain rigorous and continuous testing.
Accountability & Oversight	To ensure that AI technologies deliver their intended outcomes, human oversight must be embedded in the entire AI lifecycle, and responsibility should be allocated from the outset. It should be clear, for instance, who is responsible for the model reliability, for model review, and who bears responsibility when an AI system fails or causes harm.

By adhering to these principles organisations should be able to avoid the ethical challenges set out in [Section 2](#) above.

3.3 Socialise

Another important element of the governance of AI Ethics is socialisation. Socialisation happens both internally with the goal of creating an ethical AI culture, and externally with the goal of promoting autonomy and building trust.

a) Internal Culture Measures

The standards that an organisation has set need to be communicated to employees in various ways and on an ongoing basis. This can include awareness campaigns, but also training sessions at induction, as part of annual online and in-person training, and as targeted training for specific teams. The goal is to promote AI and data literacy and competence, so that employees are aware of the risks surrounding data and AI, and know how to use these responsibly.


Organisations should communicate the following information to internal role-players:

- *Employees*: all employees should be familiarised with the organisation's guidelines on the acceptable use of AI to ensure that AI tools are not used inappropriately by employees in their everyday work.
- *Project Teams*: the sponsors, developers and deployers of AI systems should receive training on the organisation's AI Ethics Policy as well as the AI Risk Management Standard. These employees should have a thorough understanding of the general principles that should guide their work, and the practical steps that are required as they progress along the AI lifecycle.

b) External Communication

An organisation should also communicate externally, to promote autonomy and trust. The following role-players should be provided with information:

- *Data subjects*: Whenever an organisation collects data from people, the "data subjects" need to be informed of the use of their data, how their data and their privacy are protected, how they can obtain more information about the use of their data, and how they can exercise their rights over their data. In the process, an organisation needs to obtain the informed consent of data subjects.
- *Users / Affected Parties*: Organisations also need to communicate information to the users of the AI tools deployed by that organisation, or those affected by it. This



includes finding effective ways to explain why AI tools are used, how these tools function, if and why decisions are delegated, the benefits and limitations of these tools, and channels for reporting concerns, appealing system outcomes, and for claiming redress.

Bearing the goal of human-centric and ethical AI in mind, organisations should seek the most effective ways to communicate. There is a difference between communicating for the purposes of risk management or legal protection, and communicating with the intent to inform. Organisations should strike the appropriate balance between these forms of communication. In this way organisations attend to legal requirements (compliance with rules), but also demonstrate practically the values of responsible AI, including *autonomy* and *transparency*.

3.4. Monitor & Report

Naturally, governance also involves ongoing monitoring. The committee responsible for AI Ethics governance will consider:

- The adequacy and effectiveness of measures to create an ethical AI culture within the organisation;
- Adherence by employees to the agreed standards for acceptable employee use of AI tools;
- Adherence by technical teams to the principles of responsible AI use;
- Ongoing AI system performance; and
- The adequacy and effectiveness of the control environment attached to AI use.

When an organisation buys, or develops and deploys AI systems, governance oversight will also cover the Ethics Risk Management associated with such AI projects. What this entails is explained in the next section.

3.5. AI Project Ethics Risk Management

The purchase, or development and deployment of an AI system by an organisation involves a risk management process. This type of approach is illustrated and explained below:



Project-based risk management

1. Assign responsibility
2. Consider team diversity
3. Assess against standards
4. Keep a human-in the-process
5. Set recourse avenues
6. Review regularly

a) Assign responsibility

It is crucial to recognise that the delegation of decision-making to an AI system never removes human and individual responsibility and accountability. A systemic measure for the governance of AI Ethics is therefore the establishment of clear chains of responsibility. If done correctly, it should be possible to attribute both moral and legal responsibility for every stage of the lifecycle of an AI system.

It is useful to draw a map of the different individuals and organisations involved in an AI project. This includes those responsible for the different workstreams of a project, and any outside or third-party functionaries that carry responsibility for aspects of the development or deployment of the system. There should be agreement about where responsibility lies, for instance regarding approvals, reviews and longer term (post-implementation) quality maintenance.

Those with allocated responsibility for the different aspects of the AI lifecycle are not only accountable for all outcomes of the system, but are also responsible for *transparency* and *explainability* concerning their allocated area. If needed, they may be called upon to clarify and explain the operation of the AI system (or the aspect they are responsible for).

Having clear lines of responsibility prevents the diffusion of responsibility that may lead to negligence or a situation in which the deployers of a system blame the developers for harmful outcomes, while the developers attribute blame to the deployers. Such chains of responsibility are especially pertinent where third-party organisations are involved in model design and development. The upside of clear responsibilities is that people take ownership of the system and are empowered to ensure it operates effectively and ethically.



Finally, part of this governance measure is communicating to users and other stakeholders where to report concerns and which channels to follow for redress.

b) Team diversity

Project-based Ethics Risk Management includes deliberately inviting diversity into any AI project. This is achieved in at least two ways: first, by ensuring a diverse project team; and second, by engaging with a diverse group of prospective users during the testing phase of an AI system.

Technical teams frequently lack diversity, consisting of people who share similar educational backgrounds, values and interests, as well as similar gender, ethnic and age characteristics. But it is not only the project teams that lack diversity. Often datasets and data sources also fail to reflect the diversity of the ultimate user group the AI system will affect.

The risk inherent in a lack of diversity is the possibility that important perspectives are missed, especially those of underrepresented and marginalised groups. When this happens, there is the risk that potential moral harm, bias, and discrimination remain unidentified until it is too late. The potential benefit of diversity in the project team and in the development process, on the other hand, is the early identification of bias, the mitigation of harm, the promotion of fairness, and a better-quality product.

It is therefore critical to consider a wide range of perspectives. Project teams can do this by asking questions like “Whose interests have we assumed, instead of consulted?” during risk sweeps. These questions can be answered more naturally by ensuring a diverse project team and engaging with diverse prospective users for feedback during model development.

The types of diversity one should consider include differences in gender, age, race, ethnicity, language, religion, sexual orientation, socio-economic level, and disability. But risk sweeps and testing exercises should also include different organisational perspectives. In this regard the views of Risk Management, Human Resources, Legal, Compliance, Internal Audit, and Sales and Operations.

c) Assess against standards

The diversity explained above will also assist with the effective assessment of risks and impacts related to an AI project.



Ethics risk can be defined as “choices that may cause significant harms to persons or other entities with a moral status, or are likely to spark acute moral controversy for other reasons”¹⁰. To mitigate ethics risk in the development and use of AI systems, regular *risk sweeping* is recommended – team meetings during which each team member presents potential ethics risks associated with the project.

A risk sweeping exercise should be repeated at different stages of a project, for instance during the proposal phase, the prototype phase, the testing phase, and the finalisation and quality assurance phase¹¹. Any identified risks should be assessed and documented, including the mitigating strategies decided on. These sweeps can also be used to make explicit what the benefits of the project are for individual and collective flourishing, i.e., to document ethics opportunities. As explained earlier, it is important to ensure that the team performing the risk sweeping be diverse, organisationally, but also in terms of gender, race, age, and any other relevant category.

There are many tools and checklists available to assist with AI Ethics Risk Assessments (see AI Risk Resources). Here are some of the most pertinent questions to assist in identifying AI ethics risks and opportunities:

- Who will benefit from this project?
- Who is likely to be harmed by this project?
- How did we collect the data we will be using? Was it “ethically sourced”? Were we transparent about the possible uses of the collected data? Did we provide data subjects with appropriate levels of choice in the collection and use of their data?
- Is our data accurate, complete and reliable? How long will it remain accurate and useful? Is our data diverse enough, and what could be the sources of bias in our data?
- How are we ensuring the security of the data we collected? How are we protecting the privacy of the data subjects?
- Is our project in line with our organisation’s ethical values, and the principles of responsible AI use? ([See 3.2.b above](#))

10 See Vallor, Shannon, Brian Green, and Irina Raicu. 2018. Ethics in Technology Practice: A Toolkit. The Markkula Center for Applied Ethics at Santa Clara University. <https://www.scu.edu/ethics/> [Accessed on 9 May 2025].

11 Ibid, p. 5.



- Are we using the best and most robust analytics? Have we tested and validated our models?
- Have we clearly divided and assigned responsibilities? Is it clear who is responsible for which parts of the project and the AI system? Who will ultimately be held accountable if any harms follow from the implementation of our AI system?
- Who might ultimately use this model, and how could it be misused?
- Is our AI system making any judgements that can have a significant impact on people? If so, have we embedded human oversight?
- Does a person who is affected by our AI system have a clear process for appealing the result of our system, or to report harm and claim for redress?

Many organisations already have databases or dashboards with potential AI ethics risks to consider, as well as possible mitigating actions¹². See some examples below:

Risk	Description / Examples	Mitigation Strategies
<p>Breach of Privacy Laws (including POPIA and GDPR)</p>	<p>Data used in our AI system was inappropriately and illegally collected, violating privacy laws.</p> <p>In the process of intra-organisational across national borders data sharing privacy laws were breached.</p>	<p>Obtain proper consent for collecting and processing confidential and Personally Identifiable Information.</p> <p>Involve the organisation's Information Security Officer or Information Officer from the outset of the project to ensure compliance with relevant laws.</p>

¹² Organisational AI Ethics Risk databases sometimes include more detailed Risk and Control Matrix (RACM) Self-Assessments. Such Matrices identify, categorise and assess risks while also mapping these risks against control measures. For an example of such a detailed risk matrix, see Van Vuuren, L. (ed) 2016. [Ethics Risk Handbook](#). Pretoria: The Ethics Institute.



<p>Incorrect AI system conclusions lead to harm</p>	<p>An AI system produces unreliable outcomes leading to harms and potential legal liabilities.</p>	<p>At the outset of the AI project, conduct an impact assessment to identify any possible harms.</p> <p>Perform rigorous and continuous testing of the reliability of the AI system.</p> <p>Perform a thorough bias assessment on the AI system.</p> <p>Provide users with information on the functioning of the system, and a clear process for appealing the system outcomes.</p> <p>Ensure regular and appropriate human oversight over specific outcomes and/or model reliability.</p>
<p>Breach of the EU AI Act</p>	<p>The intended use of an AI system is either prohibited or high-risk without the necessary interventions. Examples include behavioural influence (prohibited) or AI uses related to essential services (high risk).</p>	<p>Subject any AI system to a compliance assessment, including a consideration of the system's risk level and regulatory requirements associated with the risk level.</p>
<p>Loss of trust / stakeholder alienation</p>	<p>The overuse of AI systems in interacting with stakeholders leads to alienation and a loss of trust.</p>	<p>Communicate and explain the uses and benefits of AI systems to stakeholders, as well as the quality and privacy assurance measures in place.</p> <p>Provide a channel for human recourse where appropriate.</p>



AI Risk Resources

Many guides, tools and checklists are available and can assist technical teams in the process of AI ethics risk identification and assessment. Here are some examples:

Checklists and Guidelines

UNESCO. 2023. [Ethical Impact Assessment: A Tool of the Recommendation on the Ethics of Artificial Intelligence](#). Available at: [Ethical Impact Assessment: A Tool of the Recommendation on the Ethics of Artificial Intelligence | UNESCO](#) [Accessed 3 May 2025].

US AID. [Checklist for AI Deployment](#). Available at: [Artificial Intelligence Ethics Checklist](#) [Accessed 3 May 2025].

Institute and Faculty of Actuaries and Royal Statistical Society. 2019. [A Guide for Ethics Data Science](#). Available at: [An Ethical Charter for Date Science WEB FINAL. PDFA-Guide-for-Ethical-Data-Science-Final-Oct-2019.pdf](#) (See especially the Implementation Checklist at the end of the document). [Accessed on 9 May 2025].

Bias Assessment Tools

Open-source tools are available for detecting and mitigating bias in AI systems. For a list of such tools, see Kuka, V. 2024. "Essential Open-Source Tools for Bias Detection and Mitigation" Turing Post (10 March 2024). Available at: [Essential Open-Source Tools for Bias Detection and Mitigation](#). [Accessed on 9 May 2025].

A catalogue of Tools & Metrics for Trustworthy AI is also made available by the OECD at [Essential Open-Source Tools for Bias Detection and Mitigation](#) [Accessed on 3 May 2025].

d) Human-in-the-process

Closely related to the "chains of responsibility" risk management measure is ensuring human oversight of AI systems. Keeping people in the process is part of respecting and promoting autonomy, while ensuring AI systems work within defined boundaries and with limited autonomy.

The following are practical steps of maintaining human oversight and firmly establishing human autonomy in the AI process:

- Reverting to human discretion and human intervention when AI models make significant judgement or predictions (for instance the determination of fraud, the diagnosis of disease, or credit scoring).
- Reviewing and validating of AI models regularly, both planned and *ad hoc*, as self-learning AI models evolve.
- Ensuring that system decisions can always be overridden by designated team human actors.

e) Recourse

AI systems can be fallible and opaque. Such systems can therefore make incorrect determinations, or decisions that are difficult to understand. Any AI system should therefore include a clear human recourse channel. In other words, a user who is on the other side of an AI determination must be able to question or appeal such a determination in discussion with a person. This channel should be communicated and easily accessible.

Case Study: MiDAS

In 2013, the state of Michigan deployed an AI tool called MiDAS (the Michigan Integrated Data Automated System) to identify and respond to cases of unemployment benefits fraud. When the algorithm identified suspected cases of fraudulent unemployment claims, however, human oversight was not triggered, instead unemployment benefits were automatically cut, wages garnished, and tax refunds seized. The impact on those affected by these automated decisions included evictions, blacklisting, bankruptcy and homelessness.

Years later it would emerge that 85% or more of the system's fraud determinations were incorrect. Thousands of unemployed people had falsely been accused of fraud. A court settlement was reached only in January of 2024. For many of those affected, the damage cannot be undone.



f) Review regularly

Once an AI system or model has been deployed, ongoing monitoring is crucial. A system that initially produced accurate and reliable outcomes may become unreliable over time. This might be the case when datasets become outdated, when the context within which an AI system operates changes, or when issues are introduced during the model's continuous retraining (increasingly on data not only created by humans, but also by AI systems). This is sometimes referred to as "data drift", "model degradation" or "model collapse"¹³.

Continuous monitoring is therefore necessary to ensure the system is still delivering the intended output, and that it is benefiting rather than harming those affected by it.

Monitoring happens in a number of ways, and through several functions:

- *Model owners*: The sponsor or technical team that owns a model should monitor the performance of their models, regularly review the models, and consider when models need to be either updated or retired.
- *Committee Oversight*: Risk Committees or dedicated AI Oversight Committees should receive reports on the performance of AI tools at scheduled intervals to ensure proper oversight (including compliance with laws and regulation, appropriate risk management, and quality assurance).
- *Audits*: Both internal and external audit functions should perform procedures that provide assurance regarding the accuracy, reliability and performance of AI tools.
- *Customer feedback*: Through customer surveys and channels for human recourse, organisations can receive direct feedback from the users of AI systems.

13 For more information, see Marr, B. 2024. "Why AI Models Are Collapsing And What It Means For The Future Of Technology" in Forbes. Available at: <https://www.forbes.com/sites/bernardmarr/2024/08/19/why-ai-models-are-collapsing-and-what-it-means-for-the-future-of-technology/> [Accessed 9 May 2025].



C. EU AI Act - Examples

Considering the potential impact of AI on society, there is fairly little regulation and legislation. The most significant exception is the EU which passed the European Union Artificial Intelligence Act in July 2024. This is a well-considered and in-depth piece of legislation of more than 100 pages.

Our purpose is not to give an overview of this Act, as it also applies to many role-players that we are not considering for this guidebook. What might however be useful for organisations is a list of examples of different types of AI risk, and how they are categorised by the Act. This might assist organisations in their own risk assessments when designing AI interventions.



The EU AI Act significantly follows a risk-based approach to AI. The level of governance intervention that is required is dependent on the risk level of the specific AI application.

The Act gives clear descriptions and examples of what is considered unacceptable and high risk. Almost all other risk falls into the limited and minimal risk categories. In the following tables we attempt to give a simple overview of the types of risk, and how the Act specifies they be managed.

Much of the Act goes beyond these risks. For more easily accessible information see:

- <https://artificialintelligenceact.eu/>
- <https://www.twobirds.com/en/insights/2024/global/european-union-artificial-intelligence-act-guide>

Unacceptable risk

Examples:

(Note that exceptions apply that are not mentioned)

- **Social scoring by governments**
 - Classifying people by social behaviour or personal characteristics and using that to determine access to services.
- **Manipulation of behaviour**
 - Using subliminal, manipulative or deceptive techniques to influence people's behaviour and causing them harm – e.g., manipulating people into gambling.
- **Compiling a facial recognition database**
 - Especially by scraping images off the internet or CCTV cameras.
- **Biometric categorisation**
 - E.g., using AI to infer race, political opinions, sexual orientation or religion.
- **Real-time remote biometric identification in public spaces**
 - E.g., facial recognition for surveillance by law enforcement (without a legal basis).
- **Inferring emotions in workplaces or educational institutions**
 - Except for medical or safety reasons.
 - Inference of emotions during hiring is also prohibited.
 - Monitoring driver fatigue is allowed.
- **Profiling for predicting criminality**
 - Using AI to determine whether people are likely to commit a crime.
- **Exploitation of vulnerable groups**
 - E.g., manipulative advertising to children.

How to treat

- **Prohibited.**



High risk

Examples:

- **Components for use in critical infrastructure**
 - E.g., road traffic / energy grids.
- **Healthcare**
 - E.g., diagnostic AI / medical devices / drug discovery.
- **Education**
 - E.g., evaluation, assessments or application for placement.
- **Credit scoring**
- **Access to public services**
 - E.g., welfare determinations.
- **Non-banned biometric categorisation or surveillance**
- **Administration of justice and democratic processes**
 - E.g., checking the accuracy of cited laws, or applying the law to facts, or aimed at influencing voting outcomes.



How to treat

- **Compliance with specified standards and procedures**
- **Risk management**
- **Data governance (for data quality and bias prevention)**
- **Transparency (understandability)**
- **Human oversight for accountability and critical decision-making**
- **Accuracy and cybersecurity**

As you can see – many of the areas we discuss in this guidebook legally require an in-depth AI governance process in the EU. It is likely that in time this will become a standard internationally.



limited risk and minimal risk uses are not specifically listed in the Act, but the following can be limited:

Limited risk

- **Using AI for narrow procedural tasks**
 - Such as summarising meetings.
- **Improving the result of a previously completed human activity**
 - E.g., having written something and checking it for completeness or editing it.
- **Detecting decision-making patterns or deviations from prior decision-making patterns**
 - Provided that it is not meant to replace or influence the previously completed human assessment without proper human review.
- **Performing a preparatory task to an assessment that will be used for a high-risk purpose**

Examples:

- **Chatbots / virtual assistants**
- **Deepfakes**
- **Content recommendation**
 - E.g., movie / product recommendations
- **Voice assistants**
 - E.g., Siri / Alexa
- **Marketing or targeted advertising**
- **Social media filtering**



How to treat

- **Transparency**
 - telling people when they are interacting with AI



Minimal risk

Examples:

- **Video games**
- **Spam filters**
- **Email categorisation**



How to treat

- **No requirements**

South African AI regulatory framework

South Africa does not have any coherent AI regulation. The Department of Communications and Digital Technologies has however released a [National Artificial Intelligence Policy Framework](#) in August 2024.

The purpose of this framework is to guide South Africa's future policy on AI, so it is quite conceptual at this stage and gives little guidance to organisations about what is acceptable and what not.

The laws that are relevant include the Protection of Personal Information Act, the Copyright and Patent Act and the Competition Act.

Further reading: <https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-south-africa>



D. Conclusion

AI clearly comes with many advantages, but as can be seen from the above discussions, AI should not be used without controls proportionate to the risk it poses. These controls will, in many instances, have a resource and cost implication, and that needs to be factored into the business case for responsible AI use.

We are still at a very early stage of AI development, and it is important that we set the tone for responsible AI right from the start. Organisations will inevitably play a large role in how people are socialised into AI use. This guidebook aims to raise moral sensitivity to the ethical issues that might flow from AI use, as well as provide guidance on structured interventions to manage these risks. Organisations will however have to make an overt commitment to ethical AI use, as it will require effort and resources.



Some of the wisdom that we gained during our research includes the following thoughts:

- AI should not change your values as an organisation. AI should enable us to become more effective at what we do – not change who we are.
- We should not allow behaviours and impacts that were previously frowned upon, simply because AI is involved.
- AI should not lead to the decline of humans' moral reasoning ability. Ideally AI should be a tool to assist in our growth to improved moral reasoning, and moral agency.

This last point is perhaps made best by IBM in their description of Human-Centred AI.

“Human-centred AI ([HCAI](#)) is an emerging discipline intent on creating AI systems that amplify and augment rather than displace human abilities. HCAI seeks to preserve human control in a way that ensures artificial intelligence meets our needs while also operating transparently, delivering equitable outcomes, and respecting privacy.”

Ethical AI is a societal project, and we hope this guidebook is a useful resource as we take early steps in the journey.



Bibliography

60 Minutes. 2021. Yuval Noah Harari: The 2021 60 Minutes interview. [Online video]. Available at: <https://www.youtube.com/watch?v=EIVTf-C6oQo>. [Accessed on 10 May 2025].

Bird & Bird. 2025. *European Union Artificial Intelligence Act: a guide*. Available at <http://www.twobirds.com/-/media/new-website-content/pdfs/capabilities/artificial-intelligence/european-union-artificial-intelligence-act-guide.pdf>. [Accessed on 1 May 2025].

Bohannon, M. 2023. *Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions*. Available at [Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions](#). [Accessed on 4 May 2025].

Chatbase. 2025. Is ChatGPT Accurate? Latest Data & Reliability Tests (2025). Available at <https://www.chatbase.co/blog/is-chatgpt-accurate>. [Accessed on 4 May 2025].

Chelli, M. et al. 2024. “Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis” in *Journal of Medical Internet Research*. 26. e53164. 10.2196/53164. Available at: https://www.researchgate.net/publication/380789873_Hallucination_Rates_and_Reference_Accuracy_of_ChatGPT_and_Bard_for_Systematic_Reviews_Comparative_Analysis/citation/download. [Accessed on 4 May 2025].

Datatron. *Real-life Examples of Discriminating Artificial Intelligence*. Available at <https://datatron.com/real-life-examples-of-discriminating-artificial-intelligence/>. [Accessed on 4 May 2025].

Eisikovits, N. 2023. “AI Is an Existential Threat—Just Not the Way You Think”. Available at <https://theconversation.com/ai-is-an-existential-threat-just-not-the-way-you-think-207680>. [Accessed on 4 May 2025].

European Parliament. 2024. *REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 13 June 2024. European Union AI Act*. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689. [Accessed on 4 May 2025].

European Parliament. 2025. *Artificial intelligence: threats and opportunities*. Available at [Artificial intelligence: threats and opportunities | Topics | European Parliament](#). [Accessed on 1 May 2025].

Floridi, L. et al. 2018. “AI4People – An Ethical Framework for a Good AI Society:



Opportunities, Risks, Principles, and Recommendations”, in *Minds and Machines* 28, pp. 689 – 707.

Future of Life Institute. 2024. *High-level summary of the AI act*. Available at <https://artificialintelligenceact.eu/high-level-summary/>. [Accessed on 4 May 2025].

Future of Life Institute. 2024. *Pause Giant AI Experiments: An Open Letter*. Available at <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>. [Accessed on 4 May 2025].

Gillespie, N. & S. Lockey. 2025. “Major survey finds most people use AI regularly at work – but almost half admit to doing so inappropriately”, in *The Conversation*, 29 April 2025. Available at: [Major survey finds most people use AI regularly at work – but almost half admit to doing so inappropriately](#). [Accessed on 1 May 2025].

Gillespie, N., Lockey, S., Ward, T., Macdade, A., & Hassed, G. 2025. *Trust, attitudes and use of artificial intelligence: A global study 2025*. The University of Melbourne and KPMG. DOI 10.26188/28822919. [Trust, attitudes and use of artificial intelligence](#). [Accessed on 1 May 2025].

Goz, M and Spiridon, S. *AI Bias in Credit & Loan Processing: Is AI Biased When Assessing Credit Worthiness?*. Available at: [AI Bias in Credit & Loan Processing: Is AI Biased When Assessing Credit Worthiness?](#) [Accessed on 4 May 2025].


Holdsworth, J. 2023. *What is AI Bias*. IBM. Available at <https://www.ibm.com/think/topics/ai-bias>. [Accessed on 4 May 2025].

IBM. 2022. *What is human-centered AI?* Available at: <https://research.ibm.com/blog/what-is-human-centered-ai>. [Accessed on 10 May 2025].

Institute and Faculty of Actuaries and Royal Statistical Society. 2019. *A Guide for Ethics Data Science*. Available at: [An Ethical Charter for Date Science WEB FINAL.PDF](#). [Accessed on 4 May 2025].

Interaction Design Foundation. *What is Human-centered AI (HCAI)*. Available at <https://www.interaction-design.org/literature/topics/human-centered-ai?srsltid=AfmBOopqInCoTFA3rIDRc37Wq51dbRxlqWp8aJPU64ffhpdfSkShACVz>. [Accessed on 10 May 2025].

IoDSA. 2025. *King V Report on Corporate Governance for South Africa (Exposure Draft)*. Johannesburg: IoDSA. Available at https://cdn.ymaws.com/www.iodsa.co.za/resource/collection/7DAE15BF-07FA-4922-879E-6788368F5DB4/KingV_code.pdf. [Accessed on 9 May 2025].



ISO. *Building a responsible AI: How to manage the AI ethics debate.* Available at <https://www.iso.org/artificial-intelligence/responsible-ai-ethics>. [Accessed on 4 May 2025].

Kroll, JA. 2018. Data Science and Data Governance, in IEEE Security & Privacy (Volume: 16, Issue: 6, Nov-Dec 2018). Available at <https://ieeexplore.ieee.org/document/8636447>. [Accessed on 10 May 2025].

Kuka, V. 2024. “Essential Open-Source Tools for Bias Detection and Mitigation”. *Turing Post* (10 March 2024). Available at: [Essential Open-Source Tools for Bias Detection and Mitigation](#). [Accessed on 9 May 2025].

Loeppky, J. 2024. 32 times artificial intelligence got it catastrophically wrong. Available at: <https://www.livescience.com/technology/artificial-intelligence/32-times-artificial-intelligence-got-it-catastrophically-wrong>. [Accessed on 9 May 2025].

Marr, B. 2024. “Why AI Models Are Collapsing And What It Means For The Future Of Technology” in *Forbes*. Available at: <https://www.forbes.com/sites/bernard-marr/2024/08/19/why-ai-models-are-collapsing-and-what-it-means-for-the-future-of-technology/> [Accessed on 9 May 2025].

Meyer, R. 2014. “Everything We Know About Facebook’s Secret Mood-Manipulation Experiment” in *The Atlantic* (24 June 2014). Available at: [Everything We Know About Facebook’s Secret Mood-Manipulation Experiment - The Atlantic](#). [Accessed on 4 May 2025].

Mitchel, M. 2021. Podcast - Existential risk from AI: A sceptical perspective. Available at <https://medium.com/data-science/existential-risk-from-ai-a-skeptical-perspective-35f0cd7c9fa4>. [Accessed on 4 May 2025].

ProPublica. 2016. Machine bias. Available at [Machine Bias — ProPublica](#). [Accessed on 4 May 2025].

SAICA. Professional Ethical Responsibilities in the Use and Adoption of Generative Artificial Intelligence (GenAI). Available at <https://saicawebprstorage.blob.core.windows.net/uploads/Professional-ethics-in-the-use-of-Generative-AI-language.pdf>. [Accessed on 10 May 2025].

Segalla, M and Rouzies, D. 2023. The Ethics of Managing People’s Data in Harvard Business Review. Available at [The Ethics of Managing People’s Data](#). [Accessed on 4 May 2025].

UK Finance and KPMG LLP. 2019. *The Ethical Use of Customer Data in a Digital Economy*. Available at: [The Ethical Use of Data in a Digital Economy](#). [Accessed: 4 May 2025].

UNESCO. 2021. *Ethical Impact Assessment: A Tool of the Recommendation on the Ethics of Artificial Intelligence*. Available at [Ethical Impact Assessment: A Tool of the Recommendation on the Ethics of Artificial Intelligence | UNESCO](#). [Accessed on 9 May 2025].

UNESCO. 2025. *Ethics of Artificial Intelligence*. Available at [Ethics of Artificial Intelligence | UNESCO](#). [Accessed on 1 May 2025].

University of San Diego. *10 Real-Life Examples of how AI is used in Business*. Available at <https://onlinedegrees.sandiego.edu/artificial-intelligence-business/>. [Accessed on 4 May 2025].

Use of ChatGPT is indicated in the text.

Vallor, S., Green, B. and Raicu, I. 2018. *Ethics in Technology Practice: A Toolkit*. The Markkula Center for Applied Ethics at Santa Clara University. Available at: <https://www.scu.edu/ethics/>. [Accessed on 9 May 2025].

White and Case. 2024. *AI Watch: Global regulatory tracker - South Africa*. Available at <https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-south-africa>. [Accessed on 4 May 2025].

Wikipedia. *Existential risk from artificial intelligence*. Available at https://en.wikipedia.org/wiki/Existential_risk_from_artificial_intelligence. [Accessed on 4 May 2025].

Wikipedia. *Facebook – Cambridge Analytica data scandal*. Available at https://en.wikipedia.org/wiki/Facebook%E2%80%93Cambridge_Analytica_data_scandal. [Accessed 4 May 2025].

Wikipedia. *Technological unemployment*. Available at https://en.wikipedia.org/wiki/Technological_unemployment#Artificial_intelligence. [Accessed on 4 May 2025].

Yibeltal, K and Muia, W. 2023. *Facebook’s algorithms ‘supercharged’ hate speech in Ethiopia’s Tigray conflict*. BBC News. Available at [Facebook’s algorithms ‘supercharged’ hate speech in Ethiopia’s Tigray conflict](#). [Accessed 4 May 2025].

About the Authors

Kris Dobie is the Senior Manager for Organisational Ethics at The Ethics Institute (TEI), based in Pretoria, South Africa.

His involvement in organisational ethics started in 2004, and he has a special interest in corruption prevention and conflicts of interest. Over the years he has supported the South African Government with numerous policy projects, including the development of the Minimum Anti-Corruption Capacity Guidance, and the 2016 South African Local Government Anti-Corruption Strategy.

He is continuously involved in TEI's advisory and consulting activities for a diverse range of clients from the public and private sectors. He has initiated and led a number of funded initiatives, including the Local Government Ethical Leadership Initiative under which the Code for Ethical Leadership in Local Government was developed.

Kris was the lead researcher on various editions of the South African Citizens' Bribery Survey, as well as the Public Sector Ethics Survey. He authored the *Conflict of Interest Handbook*, and he co-authored both editions of the *TEI Ethics Reporting Handbook*.

He served on the Global Reporting Initiative's anti-corruption working group for the review of the ethics and anti-corruption reporting guidelines for the GRI's G4 reporting standard, and more recently served on the National Anti-Corruption Strategy Reference Group.

He holds a degree in Landscape Architecture from the University of Pretoria, as well as MPhil in Workplace Ethics (*Cum Laude*) from the same institution.

Dr Schalk Engelbrecht is an ethicist, the Chief Ethics Officer for KPMG South Africa, and a student of Philosophy. He completed his PhD in Philosophy at Stellenbosch University in 2010, with a dissertation entitled "Between Hope and Dystopia: A Critique of Utopian Reason".

He is responsible for KPMG's ethics programme in Southern Africa; a consultant on business ethics; and a research fellow of the Centre for Applied Ethics (Stellenbosch University) and the Stellenbosch Business School. He serves on the Ethics Committee of the South African Institute of Chartered Accountants (SAICA), and chairs the "Role and Mindset" working group.

Before joining KPMG, Schalk lectured Philosophy and Ethics at the University of Stellenbosch and North-West University. He has lectured Business Ethics as part of MBA programmes, and currently lectures Data Ethics for students in Data Science the University of Stellenbosch. He was previously the editor-in-chief of the *African Journal of Business Ethics*.



About The Ethics Institute

The Ethics Institute is an independent institute producing original thought leadership and offering a range of services and products related to organisational ethics.

Our vision is: *Building an ethically responsible society.*

We pursue our vision through thought leadership and an organisational ethics-related offering, including training, advisory services, assessments, products and supporter opportunities. We work with the public and private sectors and with professional associations.

All original research work produced by The Ethics Institute, including the *Ethics Handbook Series*, is freely available on our website.

www.tei.org.za

012 342 2799



The Ethics Institute



The Ethics Institute



@TheEthicsInstitute



@EthicsInst



The Ethics Institute



Guidebook to managing **The Ethics of AI in Organisations**

Kris Dobie & Schalk Engelbrecht

The fast-paced adoption of AI technologies has caused organisations across the world to ponder not only their use of AI, but also the newfound ethical risk they might be exposed to.

This guidebook is aimed specifically at organisations who are coming to grips with the ethical questions around the use of AI.

The authors have researched the field to answer these questions, and interviewed practitioners from a diverse range of organisations who have made progress in this regard. The result is a practical guidebook that looks at the ethical issues associated with the widespread use of Large Language Models by employees, as well as the governance of more specialist AI projects.

In short, the guidebook addresses the following questions:

- What are the ethical risks associated with the use of AI?; and
- What do we, as an organisation, do about it?

It therefore gives a concise and practical overview of the issues to look out for (supplemented with real-world case studies) and proposes an approach to managing the ethics of AI in organisations.

It is intended as a resource for all organisational role-players who have a responsibility to manage the ethics risks associated with AI, including governing bodies, executive teams, ethics practitioners, risk practitioners and IT development teams.