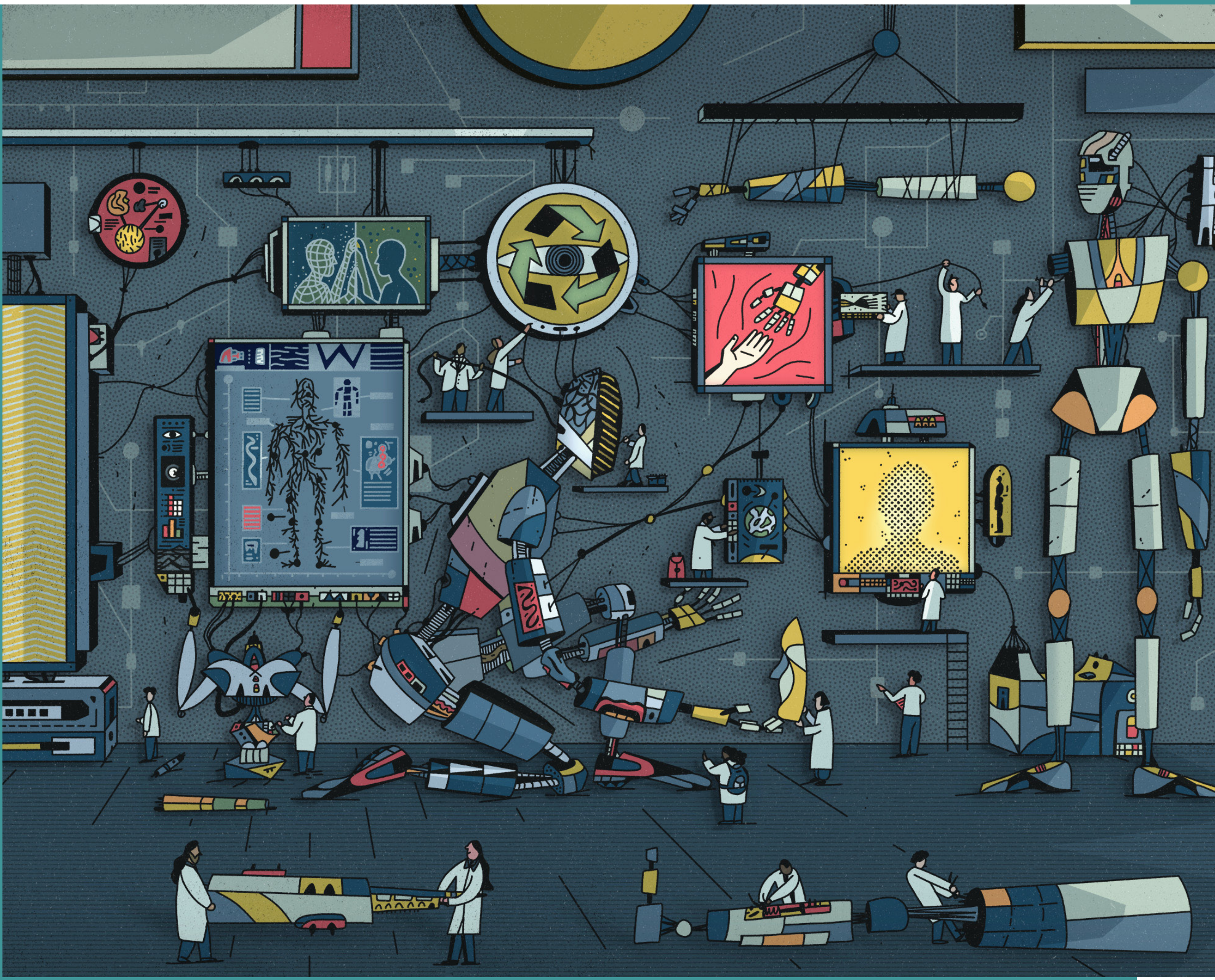
The future of trust in artificial intelligence

Responsibility, understandability and sustainability



APRIL 2022

This report proposes three hypotheses about the future of trust in artificial intelligence: a future in which trust is actively built between people and the systems they use.

It offers insights that everyone can understand and apply, from developers to designers, executives to employees and decision-makers to decision-takers.

By arguing that trustworthy technology is responsible, understandable and sustainable technology, this report explores how the risks from misuse of artificial intelligence, much like the impact of humans on our planet's climate, need to be addressed.

Contents

5 Foreword: The complexity of trust in artificial intelligence

6 Starting out: Trust and maturity

6 Understanding the risks: harms from misuse of artificial intelligence

7 Three hypotheses about the future of trust in artificial intelligence

8 **I. Responsibility is not only a role: The ethicist is necessary but not sufficient to achieve trust throughout the artificial intelligence lifecycle**

- 8 The rise of the ethicist for artificial intelligence
- 10 AI ethicist 101
- 10 "No one-size-fits-all"
- 11 And yet...
- 12 Looking ahead...
- 13 The key points

14 **II. Standardisation from diversity: Standards throughout ethical AI development can help to cultivate trust through the sharing of best practice.**

- 14 The proliferation of principles
- 14 Standard bearers: the IEEE and ISO
- 17 Calls for a coalition
- 18 And yet...
- 20 Looking ahead...
- 20 The key points

21 **III. From explainability to understandability and prospect to procedure: Technical explainability hasn't enabled trust, but a number of overlapping procedures are emerging as helpful alternatives**

- 21 Opening the black box: a lost cause?
- 22 Towards holistic explainability
- 24 Making it more understandable: the procedural toolbox
- 29 And yet...
- 30 Looking ahead...
- 31 The key points

32 Conclusions

- 32 A parallel between sustainability and technology ethics

With thanks to...

Adam Leon Smith CTO, Dragonfly

Air Cdre David Rowland Senior Responsible Officer, Ministry of Defence's AI Centre

Anna Fellander Founder, anth.ai

Anouk Ruhaak Mozilla Fellow

Beena Ammanath Executive Director, Deloitte Global AI Institute

Dama Sathianathan Partner, Bethnal Green Ventures

Dr David Leslie Director of Ethics and Responsible Innovation Research, The Alan Turing Institute

Dr Emma Ruttkamp-Bloem Chairperson, Ad hoc Expert Group UNESCO Recommendation on AI Ethics, Professor of Philosophy, University of Pretoria, AI Ethics Lead, Centre for AI Research

Dr Frens Kroeger Professor, Centre of Trust, Peace and Social Relations, Coventry University

Dr Gemma Galdón-Clavell Founder, Eticas Consulting

Gerlinde Weger AI Ethics Consultant, IEEE

Dr Jenn Wortman Vaughan Senior Principal Researcher, Microsoft Research

Kostis Manolitzas Head of Data Science, Sky Media

Lisa Talia Moretti Digital Sociologist Consultant and Associate Lecturer at Goldsmiths University

Liz Grennan Global Co-Leader of Digital Trust, McKinsey & Company

Minesh Tanna Managing Partner, Simmons and Simmons

Nell Watson Chair of ECPAIS Transparency Expert Focus Group, IEEE

Olivia Gambelin Co-founder and CEO, Ethical Intelligence

Ray Eitel-Porter Global Lead for Responsible AI, Accenture

Thomas Gray CTO and Director of Innovation, Kainos

Foreword: The complexity of trust in artificial intelligence

Alexandra Mousavizadeh, Director of Tortoise Intelligence,
and **Thomas Gray**, CTO and Director of Innovation, Kainos

Many companies are only just beginning to adopt artificial intelligence. For others, the journey is well under way. Big or small, advanced or nascent, companies at all stages of maturity face many of the same challenges, and one of those is understanding trust.

Trust is complex, especially in advanced technological systems. Artificial intelligence is currently used in a vast range of applications: from keeping spam emails out of our inboxes to sequencing human genome data, with much more in between.

When it comes to the development and use of artificial intelligence there are many factors at play. Trust in a system varies depending on the data used, the conclusions or predictions reached and the sensitivity of the system to bias and other influences. The system's history also plays a role. Crucially, artificial intelligence both influences and is influenced by the people around it.

It's a product of their decisions and biases, but also has the potential to shape those decisions and biases. All people, whether they are developers, end users, data engineers, regulators or bystanders, have a complex relationship with the artificial intelligence systems they encounter. This is a unique characteristic for a technology.

The complex, multi-layered structure of its decision-making and the specialist knowledge involved in its engineering mean that artificial intelligence is not transparent or easy to understand. If you put most of us in the cockpit of a commercial aircraft, we'd be just as useless at controlling it as we would be at overseeing a machine-learning model. Yet many of us feel we must ask very different questions about the trustworthiness of artificial intelligence than we would about other critical systems on which our daily lives depend.

Trust in technology is, in fact, interpersonal trust. Trust that people will make the right decisions. But building trust between people is not simple either. As artificial intelligence takes on more control within business, these relationships of trust will be put under new pressures. Pressures that will test the connections of trust between different parts of the wider AI ecosystem, between regulators and business, developers and managers, decision-makers and decision-takers.

We hope that this report, and the insights from its contributors, will help to advance the conversation about trustworthy artificial intelligence and give readers a sense of what the coming years might bring. We also hope it will give businesses, both large and small, a reference point for their strategies in building "human-centric" and long-lasting AI systems.

Starting out: Trust and maturity

We investigated 10 different examples of models and indices used to evaluate the maturity of a company's artificial intelligence efforts. From looking at leading companies such as IBM, Dataiku and AppliedAI, it's clear that even at different stages of maturity, companies do have some things in common.

The frameworks distinguish generally between lower, middle and higher levels of maturity.

The categories aren't completely separate, and some models use different criteria, reflecting the complex process of adoption. But we can draw from them some key ideas and attributes to help understand how maturity changes, and what that might mean for levels of trust:

- Lower level maturity – often termed "experimenters", "dabblers" or "initialisers" – is characterised by businesses that are starting out on their first AI projects, have not yet created a clear business case for further investment and are only beginning to realise value through artificial intelligence.

- Middle level maturity – often termed "explorers", "practitioners" or "expanders" – is characterised by businesses that have deployed artificial intelligence and are driving value in some way, but have not yet done so at scale.

- Higher level maturity – often termed "leaders", "shapers" or "experts" – is characterised by businesses that have created significant value through the use of artificial intelligence at scale, and continue to innovate.

Maturity varies not just between companies but also within them. Efforts to use artificial intelligence in one area of a business might be more advanced or effective than in others. But, whether businesses are large or small, mature or immature, digitised or digitising, they can all investigate *trustworthiness*.

Understanding the risks: harms from misuse of artificial intelligence

In an increasingly data-driven world, complex algorithms are being used as business solutions in a huge number of commercial domains.

"Particularly when systems are poorly designed and tested, the deployment of some solutions can lead to and perpetuate systemic harms," Peter Campbell, Data & AI Practice Director at Kainos, told us. These harms can often emerge from building on and reinforcing racial, gender or other socio-economic biases – though this shouldn't be framed as a problem inherent in artificial intelligence itself. They are down to poor design, governance and implementation. Here are some recent examples:

- Three commercial applications of AI-driven facial recognition technology were found to perpetuate racial and gender-based discrimination, as shown by Joy Buloamwini, of MIT Media Lab, who revealed the systems had low accuracy in detecting women of colour.

- Some Uber drivers were denied access to, and in some cases removed from, the company platform, harming their livelihoods, after algorithmic decision-making systems wrongly removed them because the authentication software failed to detect their faces.

- In 2017 it was revealed that Google's search algorithm was generating search results along discriminatory lines, with the algorithmic autofill suggestion for "does Islam..." adding "permit terrorism" to users' searchbars. While recommendations of a Google algorithm are a reflection of historical, social, cultural and technical factors, and although it was clearly not Google's intention to make discriminatory recommendations, producing such harmful content was still a function of the algorithmic system.

- Socio-economically discriminating algorithms were found to be used by Italian car insurance companies. A study of the opaque algorithms used to create insurance quotes found that rates varied according to citizenship and birthplace, with a driver born in Ghana being charged as much as €1,000 more than a person with an identical profile born in Milan.

It's not just in sensitive areas such as insurance, financial services and healthcare that these harms present a serious issue. Given their widespread presence, two critical questions arise:

- *How can we make sense of algorithmic systems, both in design and use, to hold people accountable for their function?*

- *How can we make sure that algorithmic systems are ethically designed, to mitigate harms before they arise?*

By addressing these two questions, of accountability and responsibility, this report shows why it's important to explore how trust in artificial intelligence is created, and how it can be sustained in the future.

Three hypotheses about the future of trust in artificial intelligence

Drawing on interviews conducted with experts at Accenture, the Alan Turing Institute, UNESCO, Sky, IEEE and other leading organisations, as well as a review of the most recent literature on ethical and responsible technology, we present three hypotheses, each shedding light on the ways that trust can be established and refined within artificial intelligence ecosystems:

I Responsibility is not only a role: The ethicist is necessary but not sufficient to achieve trust throughout the artificial intelligence lifecycle.

The rise of the ethicist for artificial intelligence

There are practical challenges to designing, developing and using trustworthy artificial intelligence. No matter what stage a company is at, it is important to ask, who do these responsibilities fall on?

For Merve Hickok, Senior Research Director of the Centre for AI and Digital Policy at the University of Michigan, there are a number of titles that are taking on this responsibility. AI ethicists are popping up as a "chief AI ethics officer, chief trust officer, ethical AI lead, AI ethics and governance or trust and safety policy adviser", to name just a few.

A sceptic might say that so many different titles, the lack of clarity of the role and the scramble to hire are signs of "ethics washing". Companies want to market themselves as trustworthy, and hiring an ethicist may be a way to do so.

Lurking beneath this trend is an assumption that ethics and trust are positively correlated: the more ethical you are, the more trustworthy you can be. Moreover, in the evolving library of reports, frameworks and principles that set out to guide responsible business practice, these terms often lose their meaning.

We need to be cautious about such assumptions. The words trust and ethics are not wholly interchangeable. Thinkers and scholars have contemplated ethics and trust for hundreds of years, and it's important to note that both concepts are fundamentally about the cultivation of good relations between humans.

Nell Watson, Chair of IEEE's ECPAIS Transparency Expert Focus Group, who is one of those thinkers in the context of artificial intelligence, and has been an executive consultant on philosophical matters for Apple, suggests that ethics are the kinds of ideas and actions that tend to drive towards greater trust.

Companies that are seeking to build trust in the people, practices and platforms that shape their AI solutions make principles of ethical artificial intelligence part of their operations.

"Moving from a *point of view* on ethics to actual operationalisation of ethics involves a big journey of understanding to know where in the workflows you'd even weigh in" Liz Grennan, Global Co-Leader of Digital Trust at McKinsey & Company told us.

The role of the AI ethicist appears to carve out a space for someone with the skills and experience to define where to weigh in, and to start doing so.

Salesforce was an early mover, announcing in 2018 that it would hire an ethics chief, Paula Goldman, with the broad remit of “developing strategies to use technology in an ethical and humane way”.

More businesses have joined in, spearheaded in the last three years by large corporations. KPMG, for example, suggested in a 2019 blog post that the AI ethicist was one of the top five AI hires that companies needed to succeed, with that role taking “the critical responsibility of establishing AI frameworks that uphold company standards and codes of ethics”.

Trust also appears to be increasingly identified as a critical aspect of emerging technology jobs. Cognizant, for example, published a report in 2018 on “getting and staying employed over the next 10 years”, in which it listed the “chief trust officer” as one of the “21 jobs of the future”, alongside positions such as “cyber city analyst”, “personal data broker” and “AI business deveIopment manager”.

A few years into that future there is little evidence that this particular role has become the dominant one. But the employment of ethicists has gained momentum across a surprisingly diverse range of sectors and AI applications, with enterprise software companies such as Hypergiant, but also more traditional retail organisations such as Ikea, making room for this sort of position.

The emergence of the AI ethicist coincides with a wave of sentiment about responsible innovation that has swept the technology sector. In 2012 a Harvard Business Review piece hailed the data scientist as perhaps the “sexiest job of the 21st century”. Today that role is much more taken for granted, though no less integral to AI implementation strategies.

While data scientists were the hot topic in the last decade, roles that steward responsible and ethical AI practices are now capturing corporate attention. In a play of words on that 2012 HBR piece, David Ryan Polgar, of All Tech is Human, has suggested that while the title of sexiest job is up for grabs by the ethicist, there is still a need to “clearly define what these new roles entail”.

> “We may want to take the AI ethicist role with a grain of salt”

David Leslie, Director of Ethics and Responsible Innovation Research at the Alan Turing Institute, told us.

> “As I see it, there are a lot of people thinking about these issues who have been for quite a while. With GDPR, impact assessments, environmental and ethical anticipation... they have been alive in the world without networks of ethicists to run these things.”

Experts in different industries are still assessing the importance of these roles. Nonetheless, there is a common set of characteristics and competencies that is giving shape to the function of the AI ethicist. Let’s look at them:

AI ethicist 101

Olivia Gambelin, the founder and CEO of Ethical Intelligence, pointed out that an AI ethicist is a person with a "critical-thinking skill set" who can identify and establish channels of cross-organisational communication. This definition chimes with a piece from David Ryan Polgar in which he outlines the ideal candidate: "someone comfortable working with both engineers and executives, typically has an advanced degree and is capable of cross-functional work regarding ethics, privacy, safety and security."

While the title of AI ethicist may end up being an umbrella term for these sub-disciplines, hiring someone who just ticks these functional boxes isn't enough. As Will Griffin from Hypergiant points out:

> "Some tech companies with a lot of money and resources hire the best tech ethicists and send them all around the world to discuss subjects like algorithmic bias, fairness and transparency. But it's fruitless, because these knowledgeable professionals don't have the buy-in to actually change the products at their own organisations. This means that all the investment put into the ethics department doesn't generate value."

Winning ethical concessions from leadership is no easy task. It can involve challenging negotiations and a re-evaluation of the priorities of the business. As Olivia Gambelin points out in a paper of hers: when acting as an ethicist in these negotiations, "you are aware that your points may jeopardise such an enormous profit" and so might be ignored. One of the key attributes of the ethicist is "bravery", she says.

"No one-size-fits-all"

Putting ethics into action is a challenge. In artificial intelligence, the work of an ethicist depends on context. Gambelin points out that "in large companies it's often a case of moving between teams and asking how decisions are made". "In smaller companies", she says, it's typically about asking: "what are the values you are designing for – values found in company policy, but also wider societal values".

Large or small, all organisations hiring an ethicist should facilitate the work they are trying to do. They need to be given the right access and the necessary power. Beena Ammanath, Executive Director of Deloitte's Global AI Institute, told us there was "no one-size-fits-all" approach to this, and that companies needed to ask: "does the ethicist have the right seat at the table? What level are they sitting at – the data science level or an executive level – to drive business process changes?"

Indeed, the level at which an ethicist operates shouldn't limit their influence. Nell Watson said the role should be cross-cutting, allowing ethicists to be an internal "nominal ethical product owner" within an AI development team who has direct communication to C-suite through something like a "red telephone".

And yet...

Although the experts we spoke to recognised the potential value of an ethicist in taking ownership of ethical practices and processes, to get ethics right for artificial intelligence requires more than just one person.

As Olivia Gambelin suggested:

> "The AI ethicist is just one piece of the AI ethics puzzle."

Indeed, an ethicist is just one element in the wider web of trust relations that are found across the development of artificial intelligence. As Beena Ammanath told us, the AI ethicist is not the "be-all solution for getting trust in AI".

Other experts also pointed to potential pitfalls in the role of the AI ethicist.

One such pitfall was described by Ray Eitel-Porter, Accenture's Global Lead for Responsible AI. He spoke about the risk of "siloing off" those who either take on the specific role as an AI ethicist or who form a dedicated AI ethics team. Specifically, some larger tech companies would allocate a role or team to ethics and the rest of the organisation then didn't worry about it because it was up to that one person or team. "You could see how that could happen, right? You have a department that focuses on this [ethics] but it is somehow isolated from the rest of the organisation."

So how can an organisation avoid these pitfalls, address the AI ethics puzzle and move towards more trustworthy relations?

Eitel-Porter explained Accenture's approach to this challenge of siloing:

> "We at Accenture very much take the view that Responsible AI is a responsibility and a business imperative that has to be embedded across the whole of the organisation and not just within the technology people. We have [ethical] training, not just for data scientists, but for everybody, essentially, who is interacting with data and AI, because we think it's everyone's responsibility to be aware of the role that they have to play, and different people have potentially different roles."

On a similar note, Lisa Talia Moretti, a digital sociologist at Goldsmiths, University of London, says that doing ethics for artificial intelligence is not a "one-stop shop": in order to effectively embed it within an organisation, a "cultural shift" is needed. This can take different forms and there are different strategies to enable it, but a key theme runs throughout: responsibility for embedding ethical practices is best diffused across an organisation rather than it landing on just one or two individuals.

At Accenture, for example, Ray Eitel-Porter's organisational responsibility is supported by what he calls a "centre of excellence". The approach here is a kind of "pull mechanism", he says, with employees being trained in responsible AI to ensure that the "leading thinking proliferates". It is then possible to check and question throughout a product lifecycle, and if an ethical dilemma emerges it can be escalated to those who have the expertise to address it.

Training is the crucial component of such cross-organisational approaches to ethical and responsible AI. Dr David Leslie, Director of Ethics and Responsible Innovation Research at the Alan Turing Institute, said the need for upskilling employees to do their technological work responsibly was the “entry cost to the knowledge economy”.

> “The knowledge economy demands a high level of upskilling in order to keep pace with the widened range of consequences that the technologies are opening up.”

Leslie’s view aligns with that of Nell Watson, who has helped to develop an online course called the Certified Ethical Emerging Technologist, suggesting that there is a clear need for “qualified individuals throughout teams who can operate a technical system ethically”.

Having multiple teams who understand how ethical dilemmas can be addressed not only facilitates trust among those employees, but it also signals to external stakeholders that AI solutions are worthy of trust because the workflows used to develop them embed ethical deliberation.

Yet training and upskilling demand time and resources. Gerlinde Weger, an AI Ethics Consultant for the IEEE and change management expert, told us that “the amount of change that’s going through companies now is colossal. If you give an employee one more thing to learn, that’s like the little wafer in Monty Python”.

One way to prevent staff from overloading while also instilling the kind of organisational responsibility for ethics, suggests Weger, is to integrate ethics into existing procedures. For example, given that “when we are talking about ethics, one of the things we are talking about is risks”, a risk-profile template, which in financial services is mandatory for any project, could embed an ethical dimension.

Weger has also found that common practices in organisational change management, such as stakeholder and impact assessments, could provide templates for questions about the ethical implications of AI development. “If you can have these [ethical considerations] as additional lenses” to procedures that are already in place, he told us, “then it’s like, oh! It flows”.

Whether a company integrates ethics into its existing workflows or decides to create new roles for this process, the work is demanding. Ultimately, however, it is an increasingly important requirement and one that communicates how organisations are addressing ethics, leading to greater levels of trust, both internal and external, in the people who steer the development of AI systems.

Looking ahead...

The role of the ethicist is an important part of establishing trustworthy artificial intelligence. It is also a potential pitfall.

The experts quoted here have shown that, while some combination of a specific role and a cross-organisational responsibility for AI ethics appears to promote trust in the architects of AI systems, there are different ways to optimise the role and ensure it can foster a widespread sense of responsibility.

This variation reflects the way in which many companies, especially at the middle and lower end of the maturity spectrum, are still finding their way through a dense and somewhat fragmented ethical AI landscape. Organisations need a steady trajectory to move from a patchwork of principles and frameworks, like those mentioned in the introduction, towards well-established best practice.

Sectors such as medicine and aviation rely on the consensus and codification of best ethical practices. In the UK, for example, the General Medical Council provides a clear set of standards from which medical practitioners make ethical judgments. In the US, the Air Line Pilots Association ensures the safety of the industry in part through its Code of Ethics.

Many experts we have spoken to see a need for a similar consensus in the artificial intelligence ecosystem on best ethical practices and how those can drive trust. It is this need that we expand on in the next hypothesis.

Roles such as head of sustainability and chief sustainability officer are now taken for granted in many organisations. Now that environmental sustainability is a key matter of public concern, the responsibility for sustainable business practices reaches beyond that individual role. Sustainability is more or less normalised, and is affecting the way businesses act on a huge scale.

Our experts suggest that ethical concerns around artificial intelligence are on a similar trajectory.

Olivia Gambelin believes that five years from now an ethicist will be employed at every company, just as data scientists are a common role today. With the seeds of ethical concern now planted, the ethicist provides the competencies to cultivate good corporate AI practice. Yet to continue growing the tree of trust – with branches of trustworthy AI architects – organisations must instil a culture that encourages all of its people to think like an ethicist.

The key points:

- The responsibility for embedding ethical practices is best diffused across an organisation rather than it landing on just one or two individuals.

- All businesses should see the competencies involved in ethical development of artificial intelligence as an “entry-level” requirement for operating in a future economy built around these technologies.

II Standardisation from diversity: Standards throughout ethical AI development will help to cultivate trust through the sharing of best practice.

The proliferation of principles

In the past five years the number of published ethical AI principles and frameworks has grown rapidly. From 2018 to 2020 there was a flood of guidelines based on the work of government initiatives, supranational bodies and private companies.

Algorithm Watch, a non-profit research and advocacy institute, found there were as many as 173 such guidelines by April 2020. Many of these documents are several hundred pages long, but it concluded that there was "a lot of virtue signalling going on and few efforts of enforcement". Out of all of these guidelines Algorithm Watch found that only 10 had "practical enforcement mechanisms".

Since that flood of guidelines, the AI ecosystem has been grappling with how businesses should actually adopt them. A somewhat tired ethical AI catchphrase of the last year on the speaking circuit is: "we need to move from principles to practice".

The thought here is that saying you are ethical doesn't justify the kind of trust in an organisation that can result from actually doing ethics.

Yet the crucial question of how to practically embed ethics into the artificial intelligence lifecycle is complex, even though one literature review suggests that fairly extensive research and development efforts from corporate and academic actors have paved the way here. In our final section of this report we will try to disentangle how some of these practices are moving from prospect to procedure.

It is important to question how the community can develop an internationally recognised certification of best practices, especially if it is to cultivate trust in what can often feel like an alphabet soup of principles and frameworks.

Many of the experts we spoke to argued that standardisation of both technical and ethical AI practices was needed.

In terms of ethical practice, which this section focuses on, standardisation is taking place across three major dimensions; certification of products and services, coalitions of ethics professionals and international standardisation of AI ethics recommendations.

Standard bearers: the IEEE and ISO

Nell Watson, Chair of IEEE's ECPAIS Transparency Expert Focus Group, is close to the leading edge on each of these dimensions. The Institute for Electrical and Electronics Engineers (IEEE) is a well-established technical professional organisation which prides itself on being a "trusted voice for engineering, computing and technology information around the globe".

Watson's work on the IEEE's newly developed CertifAIed mark seeks to engineer "credit-score-like" mechanisms, drawing on a well-defined set of ethical criteria to safeguard algorithmic trust with certification and standards.

The IEEE, Watson says, "has an amazing, long-standing pedigree and while ethics is a new element that has only arisen in the last five years or so, it's a natural outgrowth from electronic standards to practical, actionable ethical standards".

> "The need for actionable ethics standards is crucial in these times."

She adds that the IEEE mark is the only initiative that is offering this to the market.

The mark is a risk-based framework based on a number of ethical criteria. To receive it, an organisation will go through a rigorous process with an IEEE assessor, who makes an "initial assessment of the organisation, the technology and the area of operation for particular use cases to decide which criteria suite might be most useful for these particular cases".

Although the IEEE mark is still in the piloting stages, Kostis Manolitzas, Head of Data Science at Sky Media, says it has the potential to fill a gap that he feels exists across multiple sectors, even though his own, telecommunications, is "not a high-stakes industry":

> "I think we need a framework that is consolidated and initially can be a bit more generic and doesn't have to be applicable in every case, but at least it can cover the majority of the usage of these algorithms. Collaboration is going to be needed"

The IEEE is one of two leading standards bodies that are addressing the international standardisation of ethical AI practices. The International Standards Organisation (ISO) has been working on a range of standards since 2017, when, alongside the International Electrotechnical Commission (IEC), it created the expert working group SC 42 to make "headway on a groundbreaking standard that, if accepted, will offer the world a new blueprint to facilitate an AI-ready culture".

Drawing on diverse stakeholders and what the SC's Chair, Wael William Diab, calls a "management system approach", the group seeks to establish "specific controls, audit schemes and guidance that are consistent with emerging laws, regulations and stakeholder needs".

There are as many as 10 published standards under the SC 42 group, which are presented more like the "generic framework" that Manolitzas thinks will be useful. These include guidance and reference documents that outline artificial intelligence concepts and terminology, use cases, big data reference architecture and an overview of trustworthiness in artificial intelligence.

The standard on trustworthiness reads like a very initial point of reference. It provides a survey of existing approaches to improve trust, a discussion of mitigating AI risks and a discussion on improving the trustworthiness of AI systems. The extent to which it is context specific, and ready for use within different sectors remains somewhat unclear.

That said, certain implementation aspects are covered by the ISO, with a further 27 standards under development, according to their website, covering topics from machine learning explainability to an overview of ethical and societal concerns.

Some of these could provide crucial support for companies faced with a wave of regulation and legislation. As Adam Leon Smith writes for the Chartered Institute for IT (BCS), there are other ISO standards in the pipeline that include:

> "two foundational standards defining the landscape, but also the first standard that will be relevant to the legislation ISO/IEC DIS 23894: Information technology - Artificial intelligence - Risk management".

This alignment with legislation is particularly vital in light of the recent EU AI Act. The proposed law offers a four-tiered, risk-based approach to regulation. If an application is at the lower end of the risk spectrum then a lighter level of self-assessment, compliance and legal enforcement is needed, while higher risk systems will be subject to heavier, externally audited compliance requirements.

Smith told us that the price of compliance for companies developing high-risk AI products would be particularly high. He referred to work from the Centre for Data Innovation which estimated the average cost for a European SME using a high-risk system would be as much as €400,000 for the external impact assessment or audit.

The cost of compliance will also be high because of litigation, Smith said, and he thinks that larger companies will inevitably suffer less:

> "If you're a big firm and you can afford lots of lawyers to spend months agonising over it, you can probably come up with a way of complying...
>
> "Of course, litigation is not really an acceptable approach if you're an SME; you can't really be spending lots of money on technology lawyers who are very hard to come by."

Yet he also believes that technical and ethical standards can help reduce some of the financial burden of legal and regulatory compliance. The range of standards offered by the ISO in general, and its standard on AI risk management in particular, could provide a valuable barometer for anticipating new legislation and a means of minimising the cost of compliance.

But as draft legislation, the EU AI Act does not yet provide much help for companies to get ahead with their compliance. "You can't really comply with it [the EU AI Act] until you've seen the standards that are going to be written – but these haven't been written yet," Smith told us.

Calls for a coalition

So what might it take to bring about consensus on best practices and standardisation?

Collaboration can provide momentum for standard-setting initiatives such as those from the ISO or IEEE. Historically, clusters of professionals have come together towards a common ethical goal in the face of concerns about emergent harm or risk. Seán Ó hÉigeartaigh,

Director of AI: Futures and Responsibility Programme at the University of Cambridge, has called such clusters "epistemic communities" – basically a network of experts in a particular domain who seek to share knowledge.

During the Cold War, for example, "a community of arms experts helped shape cooperation between the USA and Russia... by creating an internationally shared understanding of nuclear arms control", writes Ó hÉigeartaigh.

The risks posed by AI may not be comparable to those of nuclear fallout, yet the overarching safeguarding and mitigation of risk that such a coalition might provide would clearly be beneficial.

However several experts we spoke to suggested that the community of AI ethics professionals often appears to be frustratingly siloed. In particular, Olivia Gambelin called for an international coalition of best practice among ethics specialists:

> "We are part of a coalition... of ethics firms, because we all came together and we're like, we literally all do the same process. We call it a different thing and each of us has a different spin on it. But it really is the same process. And none of us are coming together and actually collaborating because it is really a mix of people from independent firms, then you have the non profits, and then you have the people within the organisations. And we find each other and we get really excited and we exchange a few thoughts and we exchange, like, how do you approach this problem? How are you doing this?"

She adds:

> "We don't necessarily have industry standards yet. But we feel the need for it."

The coalition of ethics professionals is still nascent. But, promisingly for public trust, ethics-based accreditations are opening up for data science professionals.

In 2020, the Royal Statistical Society (RSS) indicated that it would be working with a number of other professional associations to develop industry standards for data science professionals, to "ensure an ethical and well-governed approach so the public can have confidence in how their data is being used". Moreover, the newly launched Association of Data Scientists now offers a Chartered Data Scientist qualification, which the association describes as the "highest distinction in the data science profession".

Much like the trust promoted by a hybrid approach to AI ethics that combines the role of the ethicist with a cross-organisational responsibility, these certification, standardisation and coalition developments have the potential to build a bedrock of public trust in those shaping the artificial intelligence ecosystem.

And yet...

While accreditation, certification and professional coalitions may help bind relationships of trust, several of the experts we spoke to believe that for standardisation to be successful at a global level, the community needs to find a way to reach an international consensus from a broad diversity of opinions and approaches. This is no mean feat.

There is need for diversity and a sensitivity to context.

> "For AI ethics, there isn't necessarily one approach – there is a need for a diversity of perspectives."

This comment by Ray Eitel-Porter is at the heart of the conversation on standardisation. On the one hand, as David Leslie suggested, with ethics, trust and AI "we are facing universal problems and so there is a need for global-level recommendations". Yet on the other hand, what is ethical and what enables trust can vary according to different cultural values. Leslie reminds us that we need to think about how we account for Ubuntu or other relational value systems that millions of people the world over believe to have meaning.

The question for the ethical AI community is how can a diversity of perspectives be included to reach universally actionable international standards?

One obstacle stems from a point made by Leslie: that the "global AI innovation ecosystem is dominated by the Global North" – the conversation on both ethics and trust is overshadowed by North American and European players.

This has led to what Emma Ruttkamp-Bloem, Chairperson for UNESCO's Recommendation on AI Ethics, called an "epistemic injustice" in the Global South. When we spoke to her, she was passionate about the need to make universal ethical principles work in local contexts.

So how can this challenge be addressed? David Leslie has recently been working as the lead researcher on a project, PATH AI, which is exploring "different intercultural interpretations of values such as privacy, trust and agency" and how those "can be applied to new governance frameworks around AI".

Although the project is only at the research and consultation phase, it seeks to shape "the international landscape for AI ethics, governance and regulation in a more inclusive and globally representative way".

What PATH AI sets out to do echoes what UNESCO has done on policy with its Recommendation on AI Ethics. This brought together philosophers, lawyers, diplomats, practitioners and UNESCO's Secretariat in a group of 24 people to represent the six UNESCO regions. There were consultations for each region and with young people. The results were revised after input from intergovernmental institutions then submitted to member states for a "landmark" diplomatic negotiation, Emma Ruttkamp-Bloem told us.

Such an agreement is needed, she says, because of "the multinational status of the big tech companies, meaning that international laws and principles are key". But perhaps what it demonstrates is how a shared understanding of ethical best practice can be established through an internationally representative and culturally sensitive process. It's proof, Ruttkamp-Bloem says, "that this kind of collaboration is possible... and it helps countries in the Global South to at least have something to guide them".

The Recommendation provides a rigorous set of principles which can then be applied in particular languages and cultures. Yet Ruttkamp-Bloem makes it clear that protecting everyone equally from harm is a key objective.

On the principle of privacy, for example, she observes that although it may be against local values in African communities, in which a collectivist culture and ethic may make people more open about their private lives, this does not mean those people do not have a fundamental right to privacy. "There has to be sensitivity to cultures", she says, "but there has to be the same protection against harms for everyone." This is what the UNESCO Recommendation provides.

Balancing universal rights with specific cultural values may seem like a massive task for private sector enterprises, especially those who are just starting out on the artificial intelligence journey. Yet Ruttkamp-Bloem has a few suggestions about how commercial organisations can realise business value through policy:

> "In the end it is a strategic move to be ethically sensitive and culturally aware. You can do this by ensuring your tech team is diverse to help naturally detect biases, and respect your clients and their demands and rights by ensuring transparency."

She adds that a key ethical mindset that companies can adopt is to try and "meet people where they are". Beena Ammanath raised a similar point: "We have to get down from this top level and get down in the weeds to figure out how this is going to happen – that's part of the operationalisation, making it context-specific."

She described how this tension between universal standards and context specificity could play out:

> "With regulations and standards and best practices there is a reason we have different ones for different industries... We will start with broad ones, and then there's also movement going on from the bottom up – so we will have specific content for specific application areas."

For Ammanath, it makes sense that movements towards standardisation and consensus start with broad reference points such as the UNESCO Recommendation, but she also thinks sector-specific standards will emerge from the bottom up.

When we asked her whether she thought the EU AI Act would have content and sector-specific application areas, she replied: "Absolutely", telling us that standards would inevitably take a similar approach.

The approach, then, Ammanath says, is “you start from a broad overarching umbrella of do-no-harm, and then you ask ‘how do you actually make it real within a bank versus within a hospital?’”

Looking ahead...

In sustainability, global standards have been around for at least 30 years – from ecolabels and organic food labels to social welfare standards that aim to protect workers in ‘sweatshop’ factories. Today there are as many as 400 established standards and certifications, according to one NGO that demonstrate the sustainability performance of organisations or products in specific areas.

Evidence suggests such standardisation is further off for artificial intelligence, with the development of a technical standards hub for AI having only been recently announced by the UK government.

Yet our research has also led us to believe that a mature set of actionable standards and benchmarkable certifications, such as those of the IEEE, provides a helpful guide for best ethical practices in the development and deployment of artificial intelligence.

Even though the path to standardisation still seems uncoordinated and the uptake of standards and certifications is still in the pipeline, a wave of regulation is breaking.

The EU AI Act provides a codification of the trustworthy AI paradigm which, according to Mauritz Kop of Stanford Law School, “requires AI to be legally, ethically and technically robust while respecting democratic values, human rights and the rule of law”.

The cost of compliance with what seems like imminent AI legislation will be financially unsustainable for many companies; to keep pace with the changes, however, standardisation can reduce the burden.

While regulations don’t fully mitigate ethical risks and harms, and although legal enforcement has not yet arrived, the set of proposed standards and certifications outlined in this section will give organisations a means of sense-checking their AI practices. It is likely that some of these, such as those developed by the ISO IEC SC-42, will become the gold standard that reflects the EU AI Act.

For now, what might seem like quite generic standards, such as those from the ISO covering trustworthiness or risk management, will at least help companies to align with initial regulatory moves such as the EU’s trustworthy AI paradigm and to anticipate later waves of harder legislation to come.

The key points:

- The time has come for actionable ethics standards on artificial intelligence; there are a few organisations leading the development of such standards.
- Businesses have the opportunity now to prepare for an imminent wave of legislation.
- There is a business advantage to anticipating these developments while also being mindful of culturally diverse perspectives and inclusive ethical practices.

III From explainability to understandability and prospect to procedure: Technical explainability hasn't enabled trust, but a number of overlapping procedures are emerging as helpful alternatives

Opening the black box: a lost cause?

Commercial artificial intelligence relies on complex processes. Whether it's advanced statistical modelling, machine or deep learning techniques, decisions are being made using very complicated maths.

In a blog post, Faculty, a leading AI solutions provider, suggests that there is often a correlation between complexity – the inner working of the so-called algorithmic "black box" – and the performance of the system: the more complex and seemingly unexplainable the behaviour of the model, the more accurate it can often be.

There's widespread acknowledgement that poorly designed decision-making systems can lead to unsafe and potentially harmful machine behaviour, often stemming from human biases that may reflect racial, gendered or socio-economic differences. Observatories, like the one maintained by the Eticas Foundation, can help us to trace these harms.

Given their mathematical complexity and societal importance, there have been many efforts to explain algorithmic decision-making systems. Academics, industry practitioners, government organisations and NGOs have tried to explain the behaviour of such systems. This is known as explainability. The logic of explainability is that the behaviour of algorithmic decision-making systems should be justified and subject to scrutiny so that the societal impacts can be more successfully managed by both businesses and governments.

Calls for explainability have been spearheaded by global technology companies and academic researchers through the development of technical methodologies and toolkits that are often grouped into a sub-discipline known as explainable AI (XAI).

XAI emerged around 2017 to help AI practitioners explain complex model behaviour to other practitioners: IBM has an AI Explainability 360 Kit which provides eight "state of the art explainability algorithms add transparency throughout AI systems"; Google has developed Explainable AI to provide "human-interpretable explanations of machine learning models through its tools like "AutoML" and "Vertex AI"; and Microsoft has its Interpret ML tool to analyse models and explain behaviours.

Yet, while these explainability techniques profess to "grow end-user trust and improve transparency", as Google claims, the link between giving a technical explanation and creating trust is not always clear.

Jenn Wortman Vaughan, a Senior Principal Researcher at Microsoft Research, told us:

> "Providing explanations can cause data scientists – as relative experts – to overtrust machine-learning models. What we should be aiming for is appropriate trust: how can we boost stakeholders' trust in the system when it is doing the right thing, while fostering an understanding of the limitations of the system and what can potentially go wrong?"

Similarly, trust expert and Research Professor at the University of Coventry, Dr Frens Kroeger, told us that for most stakeholders in the artificial intelligence ecosystem, the technical tools of explainability don't necessarily provide appropriate or well-placed trust.

Talking about the way that explainability has developed, Kroeger suggested that much of it "was just purely hard technical explanations... where software engineers interpret them and then you say, 'yeah, okay — that's trustworthy'".

> "But the explanation doesn't make sense to a vast majority of the people out there. Because they're not experts."

The sense we got from Kroeger was that there is a need to go beyond just technical explainability tools if we want to encourage well-placed trust in AI systems:

> "Can we get away from technical explanations and can we try and devise social explanations?"

Instead, he prefers explanations that reflect "the sort of institutional framework that surrounds the development of artificial intelligence; that is:

> "How can we develop an explanation that can in some way make the values of those companies that are behind it a bit more tangible?"

Kroeger calls it expanded or social explainability, a phrase that puts the focus on explaining the social contexts and values that surround algorithm development and deployment decisions. It's an idea that echoes two of the four pillars the ICO and the Alan Turing Institute gave for their guidance on explainability, notably, "consider context and reflect on impacts".

Towards holistic explainability

Following a similar line of thought, David Leslie told us about his work as lead researcher on the joint ICO and Alan Turing initiative Project Explain. He talked us through what he calls a "topology of explanations" to "do explainability from a more holistic point of view".

Beyond what Leslie calls the "rationale explanation" (the technical component that explains the function of an algorithm), the multitude of approaches within this holistic explainability include:

- **Impact explanations:** "Have you built in mechanisms for making sure that impacted stakeholders will be privy to explanations of how the ethics has been done and what decisions have been made and the deliberation behind those kinds of choices?"
- **Data explanations:** "Being clear about the provenance of the data set."
- **Fairness explanations:** "Being able to demonstrate across the design, development, deployment lifecycle that a project team has sufficiently considered potential pitfalls of bias, and that there has been a deliberate and transparent approach to defining what fairness criteria are being incorporated into the system."
- **Responsibility explanations:** "Being transparent about who, at any given point across the lifecycle, owns and is involved in the decision-making"

So beyond the technical explainability, **"there's all these other kinds of needed explanations that can justify public trust in the system"**.

This speaks to the need for context-specific transparency highlighted by Jenn Wortman Vaughan:

> "There are different stakeholders of AI systems who require transparency, including data scientists building systems, regulators, decision-makers using AI systems, and those who are ultimately impacted by the systems. These stakeholders have different goals, different expertise, and therefore different needs, so the approach that works best for one may not be helpful for another."

She details some of the different stakeholders and the transparency they need:

- **For technical practitioners:** "If we're thinking about a data scientist trying to debug a machine-learning model, they might benefit most from a tool like InterpretML which provides [both specific and general] explanations of the model's behaviour."
- **For business stakeholders:** "A decision-maker who is trying to determine whether or not an AI system is appropriate for their company may be better off with a clear description of the intended use cases, capabilities and, perhaps most crucially, limitations of that system. This is what we designed Microsoft's Transparency Notes for."
- **For regulators or compliance officers:** "What's needed may be an understanding of the characteristics of the particular dataset that was used to train a model, in which case a datasheet may be most appropriate."

A thread that runs through the work of Leslie, Kroeger and Wortman-Vaughan is a focus on the people and contexts surrounding AI development – what some experts have called "meaningful transparency". Meaningful transparency, as a blog post from the Ada Lovelace Institute notes, is what gets us to a place of "genuine understanding and engagement with algorithmic processes" by pulling the social and policy dimensions of AI development decisions into focus.

Meaningful or contextual transparency, holistic or social explainability – some may think it's all just semantics. But there is a common theme here: **engaging the people and contexts beyond the technical components of a system can promote understanding of, and well-placed trust in, artificial intelligence.**

Making it more understandable: the procedural toolbox

Many of the experts we spoke to have been involved in creating procedures to promote well-placed trust in AI architects and products. They have done so in ways that echo Leslie's holistic characterisation of explainability. Some of these procedures are still evolving, and there was a consensus that there isn't one silver bullet to solve the problem of trust, but many complementary approaches.

Nonetheless, being familiar with these developments now is an advantage to any organisation developing trustworthy AI practices, depending on where they are on the maturity spectrum.

Decision Documenting

Lisa Talia Moretti, for example, suggested that a process of decision documenting was crucial to enable the social transparency of AI workflows. This might be:

> "a single user researcher or a single anthropologist working within the team whose sole job it is to actually document the way that the team actually went about this and document that decision-making process. Or you could do something more collectively where you have a team who maintain weekly notes and have a constant trail of meetings around an AI product decision; noting in a few lines that this is who we spoke to, these are the decisions we made."

Sharing how decisions are made can be a key driver of trust internally and externally, according to a project led by the Partnership on AI called ABOUT ML. A blog post from ABOUT ML suggests that "the process of documentation can support the goal of transparency by prompting critical thinking about the ethical implications of each step in the artificial intelligence lifecycle and ensuring that important steps are not skipped".

But how documentation is used and what explanations it provides can vary according to who it is used for, as Beena Ammanath told us:

> "Each stakeholder requires different levels of explanation for the AI solution."

One sector in which explaining algorithmic decisions is a crucial responsibility is in defence applications.

We spoke to Air Cdre David Rowland, the Senior Responsible Officer for the Ministry of Defence's new AI Centre, who told us that although the MoD currently doesn't use artificial intelligence in high-stakes contexts, the need for explaining decisions is massive "because of the nature of what defence actually does":

> "A lot of what defence does is deterrence and sub-threshold cyber type activity to show those that could potentially do us harm how strong we are in that environment – AI will play a part in that future."
>
> "Some of it is just to show them that they can't attack our network so they can't commit fraud against us, or for our workforce not to jump on the wrong links...
>
> "But of course if that all goes wrong then we do need to create violence and harm against those that would do us harm...
>
> "Therefore, it's absolutely incumbent upon us that we make sure that if we have got life and death decisions, then we absolutely understand the mechanisms involved in those decisions."

Processes of decision documenting clearly have a crucial place in many levels of ethical and safety-related risk.

Explainability statements

Where the process of decision documentation is used to report internally on how technical and strategic choices are made in a system, an explainability statement can help inform external users about the AI used in algorithmically supported platforms.

Minesh Tanna, Managing Associate at Simmons & Simmons, introduced us to an explainability statement – the first of its kind in the world – that he worked on for a health management and self-care app called Healthily. The statement provides users with information on "how the artificial intelligence in our app works, including when, how and why we use this technology". HireVue, the leading HR interview platform, has since followed on with a similar document, also reviewed by the UK's Information Commissioner's Office under the aegis of their implementation of the GDPR's rules on automated decisions.

Tim Gordon, co-founder of Best Practice AI, an AI Management Consultancy who partnered with Simmons & Simmons on the Healthily work, suggested in a recent blog post that explainability statements might require some work – providing transparency on data sourcing and tagging, and showing how algorithms are trained and what processes are in place to respond to and manage harms is not necessarily a simple task.

But he sees five reasons why businesses should consider preparing one:

1. **The legal expectation:** In Europe, under GDPR, you need to be able to explain how, where and why you are using AI.
2. **Internal organisation:** It brings stakeholders together to make sure nothing "slips between the cracks".

3. **Customer value:** It provides detailed information for those who want to know about how algorithms are used. An AI Explainability Statement provides the material to generate even one-page summaries - as for example HireVue has done from their work with Best Practice AI.

4. **Limits liability:** Court cases in Europe, such as those against Uber/Ola in the Netherlands, have set clear precedents that you need to explain what is going on if AI is affecting individual workers.

5. **Growing international expectation:** With regulation in China, New York and California moving in the direction of transparency, explainability statements are globally relevant.

He argues that ultimately the investment in transparency is the path to generating trust.

One-pagers, leaflets and Nutri-Score labels

A 13-page explainability statement may not seem that long compared with the several hundred pages of AI ethics principles that some organisations have published. Yet to an end-user with a relatively low – even non-existent – knowledge of artificial intelligence, it may be optimistic to think such a document can offer much value.

At the other end of the spectrum, the one-page toolkit developed by Rolls-Royce, known as the Aletheia Framework, provides a practical guide for developers, executives and boards before and during usage of artificial intelligence. Rolls-Royce's Head of Service Integrity, Lee Glazier, who spearheaded the development of the Aletheia Framework, said it was a response to a wave of long and impractical ethical frameworks that emerged just over two years ago.

Instead they created an "A3 sheet of paper, that is really agile and developers can fill it out really quickly", he explained.

Going beyond understanding an AI system, the framework asks those developing and deploying AI to consider "32 facets of social impact, governance and trust and transparency and to provide evidence which can then be used to engage with approvers, stakeholders or auditors".

In terms of explaining the algorithmic behaviour of a system, the framework adopts what Caroline Gorski, the Group Director at Rolls-Royce's R² Data Labs, calls an "input-output assurance model": "It ensures the inputs are trustworthy and outputs are trustworthy; it does not require you to publish every element of the model in between."

Gorski's reasoning for this approach echoes the concern that was relayed over and over again during the interviews we conducted:

> "It is profoundly difficult to explain those black box models. While there is lots of good work on this, in our view it is probably several years, if not a decade, away from being possible."

This type of document has practical value, and the framework is now available open-source following interest from organisations in many other sectors.

It has made an impression, Glazier says. "It was deemed, even by big tech and our peers, as something that was unique because it was accessible and it was an A3 sheet of paper, rather than a 100-page document."

We saw a similar practicality in the "algorithmic leaflet". The team at Eticas Consulting, led by Gemma Galdón-Clavell, is developing the leaflet with several governments for use alongside their labour algorithms.

> "Taking the idea of the medical leaflet for when you buy medicine: it comes with a document that tells you how to consume that medicine in conditions of safety. It tells you about the ingredients that go into it and you don't always understand everything. But it's a document that helps regulators and the public understand what some of the impacts of that piece of medicine are."

For an even more user-friendly form of explanation, her team is working on Nutri-Score labels to provide a "visually comprehensive way of understanding" and "comparing between different products offered in algorithmic decision-making... like in Europe where you have an ABCD system for kitchen appliances".

Algorithmic impact assessments, audits and assurance

Being "historiographers" about ethics and trusting artificial intelligence, says David Leslie, it's clear "there have been various waves, from principles to practice to building assurance mechanisms".

The mechanisms discussed so far in this section provide a few options to build multi-stakeholder trust that go beyond the first stages of this ethical AI journey, by focusing on the understandability and social transparency of AI development and deployment.

Yet to help reach a critical mass of public trust across the AI ecosystem, many of the experts we spoke to are working on initiatives that provide assessment, reporting and assurance of ethical and responsible AI practices and impacts.

Understanding a system and the decisions that shape that system is one side of what you could call the "trust-through-transparency" coin. On the other side of that coin there is the need to hold those involved with the development and deployment of that system accountable for those processes and decisions.

One "emerging mechanism" to build algorithmic accountability, says the Ada Lovelace Institute in a recent report, is the Algorithmic Impact Assessment (AIA). In a recent report from Data and Society, the authors suggest that impact assessments offer a "means to describe, measure and assign responsibility for impacts without the need to encode explicit scientific understandings in law." They go on to suggest that the widespread interest in AIAs "comes from how they integrate measurement and responsibility". Drawing on both sides of our trust-through-transparency coin, "an impact assessment bundles together an account of what this system does and who should remedy its problems".

But as the Ada Lovelace also notes in its report, “AIAs are not a complete solution for accountability on their own: they are best complemented by other algorithmic accountability initiatives, such as audits or transparency registers.”

Audits are increasingly talked about among ethical AI experts. Although by no means at a stage of mainstream adoption, they are being proposed as the next step in the trust trajectory. Why? The thinking is that they could provide assessment and reporting processes to make algorithmic systems more accountable, creating a rigorous ecosystem for assuring that systems are developed in ways that are deserving of trust.

Although audits are frequently proposed using a comprehensive methodology such as that developed by Gemma Galdón-Clavell at Eticas Consulting, there are some existing tools, such as IBM’s AI Factsheets and Google’s Model Cards for Model Reporting, that companies could integrate into an AI workflow to provide the documentation and reporting that would support an audit.

Ultimately, the overarching goal of an auditing process, says Emre Kazim of Holistic AI, is to “improve confidence in the algorithm, ensure trustworthiness of the underlying system and convey both with a certification process”.

And although there don’t appear to be any concrete examples of what that certification will look like, the idea that audits can enable trust was underlined by Galdón-Clavell, who has been doing audits now for three years:

> “Algorithmic audits are one of the most practical things we can do in terms of increasing trust in AI... it’s about taking back control.”

The audit can provide a practical component to the kind of AI assurance ecosystem that is mapped out in a recent report from the UK government’s Centre for Data Ethics and Innovation (CDEI). The CDEI suggests that AI assurance services, such as certain forms of audit, will help to verify independently the trustworthiness of AI systems; trust here hinges in part on the quality of the system and whether companies “do what they claim in the way they claim”.

The CDEI sees assurances playing a “critical role in building and maintaining trust as artificial intelligence becomes increasingly adopted across the economy”. Indeed, the report optimistically suggests that an industry may grow around assurance comparable in size to the market for cybersecurity, which generated some £4 billion for the UK economy in 2019 and employs 43,000 people.

In terms of regulation, experts have pointed to the audit as a critical measure to achieve compliance with legislation that appears to be on the horizon following the EU AI Act. Yet, as Adam Leon Smith’s comments from the previous section suggest, the cost of compliance could be massive, with audits involving expensive preparatory work before an auditor even comes in.

Yet it's a necessary cost, he adds. Just as with other industries such as financial services, this is what doing business often entails.

> "[If] I'm making a change to critical banking infrastructure, I have to be ready for an audit as well. You know, I have to have huge amounts of evidence of what I've done and why I've done it. And that's the cost of doing business in that space. I think because of the risk introduced by AI, this is just part of the cost of doing business."

And yet...

There is some debate about whether the audit actually does enough in terms of addressing the risks and harms that stem from poorly designed artificial intelligence. As David Leslie warned:

> "I don't think we are there... and there is a risk that audits lead to a surface-level demonstration that doesn't address underlying problems."

For Leslie, such surface-level mechanisms are born from what some call an 'ex-post' approach to technological innovation – that is, where mechanisms that address the trustworthiness of a system are drawn on after the development itself has taken place.

The danger seems to be that no matter how advanced certain mechanisms for disclosure and assurance are, they can often be taken as an end in themselves, rather than a way of reflecting an ongoing compliance and deep-rooted alignment with ethical principles.

A sceptic might point out that the increasingly mature set of environmental sustainability disclosure frameworks, such as those from CDP, don't address the underlying issues of carbon intensive business and emissions. According to this view, despite the looming disaster of climate change, businesses haven't moved far or fast enough to address them.

Gemma Galdón-Clavell, whose work has largely been focused on conducting algorithmic audits, acknowledges the criticism of audits that don't go far enough. Yet she also told us that the audit methodology of Eticas emphasised an "end-to-end consideration", looking "not just at the technical properties of a system, but the issues of power redistribution and social inefficiencies that are found alongside it".

The audit process the Eticas team carries out often has the potential to change how teams understand their algorithms and the risks associated with them, rather than simply reflecting the characteristics of the algorithms themselves.

> "When we get in, everything is disorganised: often no one knows where all the data comes from. Whether it's legal to use it, what it's doing, no one's defined the protected groups. So we need to do a lot of work together and that becomes a learning process for them."

The audit, as a consultation process, is more than just a tick-box compliance exercise. Galdón-Clavell feels it can instil a sense of ethical awareness and responsibility in a team. “We are changing that team and we are changing the view of organisational responsibility and awareness around those issues and the role that they play.”

Eticas is one of the only organisations actually doing algorithmic audits right now. Peter Campbell of Kainos points out that such a rigorous, externally mediated approach to identifying, reporting and assessing ethical harms may end up being a challenge to scale.

Nonetheless, the audit provides perhaps the most comprehensive example of “trust through transparency”, and there is reason to think it will be a key enabler of well-placed trust among stakeholders within and beyond an organisation developing an algorithmic system.

Looking ahead...

The documenting and disclosing of harms and risks – often known as negative externalities – is an established financial reporting practice across almost every sector. What is more, frameworks for both internal and external auditing and assurance that hold organisations accountable are nothing new.

Environmental harms, particularly those relating to carbon emissions, are becoming a more central part of financial reporting. Indeed, the International Sustainability Standards Board (ISSB) has recently become one of two pillars for the International Financial Reporting Standards (IFRS) to provide “investors with transparent and reliable information about a company’s financial position and performance, as well as information about sustainability factors that could create or erode its enterprise value in the short, medium and long term”.

For example, KPMG suggests that there is increasing pressure from investors and consumers to coordinate financial and non-financial statements about climate-related impacts. That is, there is a growing need to show the link between statements that are found at the top of annual reports – where strategic corporate decisions such as “net-zero targets” might be documented – and the assumptions that inform the financial statements further down.

There are even moves to integrate negative environmental externalities into financial modelling in the form of impact weighted accounts.

This movement towards a forward-looking integration appears to complement the kind of corporate sustainability envisaged by David Leslie; anticipating potential harms and risks by building contingencies into models from the start is a good way to implement sustainable practices from the outset.

On artificial intelligence we are yet to see such progress on reporting ethical risks and harms, let alone its integration into corporate financial accounts and strategy. Yet corporations can take note: **it is possible that ethical and trustworthy AI will follow a similar path of integration into financial reporting to that being taken by environmental sustainability.**

Even before the artificial intelligence ecosystem reaches that point, the evolution of reporting and documenting procedures is promising, from actionable processes such as decision documenting and explainability statements to more comprehensive and wide-reaching methods such as algorithmic audits and assurance framework. These processes can foster a similar culture of transparency and accountability to that seen in environmental sustainability. This bodes well for the trustworthiness of artificial intelligence at scale.

Crucially, as the use of high-risk and complex algorithms continues to increase, and transparency-enforcing legislation materialises, the need for the end-to-end auditing support that Gemma Galdón-Clavell and her team provides will become more and more important. As she tells us:

> “I think that in the next five years, if we take into account all the trends and add in the avalanche of legislation at the EU level, I think that there’s going to be a lot of change....
>
> “Does that mean that in five years we’ll be able to just audit as a financial auditor would? Just go in and verify that things work, and then get out? That may take a bit longer. But I think that significant change is pretty much around the corner.”

What kind of change might that be? Well for the UK government’s CDEI, that change is manifested in its vision of auditing and assuring artificial intelligence – and the message is clear: “The UK will have a thriving and effective AI assurance ecosystem within the next five years.”

The key points:

- Engaging beyond the merely technical components of a system enables understanding of artificial intelligence, and can encourage well-placed trust in it.
- The deployment of AI involves a range of ethical risks. Particularly in settings where those risks are high it is critical to understand not only how the system behaves but also how decisions that shape the system are made, and what mechanisms are involved in those decisions.
- It is possible that ethical artificial intelligence will follow a similar path of integration into financial reporting to that being taken by environmental sustainability.

Conclusions

A parallel between sustainability and technology ethics

In the past two decades, sustainability has moved from being a less recognised aspect of business to perhaps the most dominant commercial narrative in history.

Where, before, chief sustainability officers were rare, now there's barely a large corporation on the planet that doesn't have one.

The more investors and consumers have become critical of the effects of a carbon-intensive "business as usual" approach, the more companies have responded with environmental sustainability initiatives, hires and commitments.

Although the impacts of artificial intelligence are arguably lower down on the list of critical societal challenges, some lessons from corporate sustainability movements can help to prevent the challenge of trust in artificial intelligence from escalating to the state of crisis we have seen with climate change.

"ESG will grow to include the artificial intelligence vertical and data governance, because they are core governance issues." Liz Grennan believes; having seen that algorithmic impact assessments, privacy programs and data officers are becoming necessary around the world.

Specifically, trust in artificial intelligence can be supported by the same sort of actions that businesses are using to demonstrate their environmental sustainability within the ESG frame: the mechanisms, commitments and strategies discussed in this report.

For David Leslie, the challenge is clearly present for both environmental sustainability and trustworthy artificial intelligence. What is needed, he contends, is a shift in focus towards real democratic governance:

> "Why are companies still polluting the environment? Because there's a higher level of accountability to the boards and to optimising profitability than there is to real democratic governance of the corporate practices, and the same will be true in the artificial intelligence ecosystem. This is to say that the more that you have inclusive involvement of impacted stakeholders in the decision-making mechanism, true democratic governance, the more transformation you will see in the practices themselves."

Leslie thinks that we are at a formative stage in the way we design and govern artificial intelligence systems. We have yet to see an 'ex ante' mindset emerge – that is, not thinking about how we can make existing artificial intelligence systems more trustworthy, but how we embed ethical design practices into teams and products from the outset to achieve greater trust in those systems. This mindset, he says, is ultimately going to come from a "shift of culture towards the democratic governance of technology".

The three hypotheses: a call to action

1. Responsibility is not only a role

The ethicist, despite being a pivotal steward of ethical artificial intelligence, shouldn't be seen as the only solution to establishing trust.

Much like the effect that public concern about climate change has had on environmental sustainability practices, if the artificial intelligence ecosystem is to reach a point where an awareness of ethical harms can exist within all development practices, an organisational culture in which everyone has an ethicist's perspective seems crucial. The responsibility can't remain siloed within the remit of an individual.

Ultimately the role of the ethicist is likely to become more and more taken for granted as the procedures become normalised. Nonetheless, if we are to realise the full potential of artificial intelligence to transform the world for the better, the responsibility (not just the role) of ethics needs to become more and more embedded.

Companies can begin by seeking out ethics professionals and giving them the remit to change the way the responsibility for ethical development is communicated and addressed. Executive and board level sponsorship of ethics initiatives is also crucial. This makes the subject of ethical application much more visible, and puts it on the same footing as other key organisational objectives. Ask who is accountable for driving this initiative forward, and how they will be quantifying and reporting on progress; many of the mechanisms discussed in Chapter 3 of this report are not only ready for adoption right now, but also provide measurable outputs.

2. Standardisation from diversity

While there have been tentative steps towards international standardisation and certification, there is insufficient coordination between the many movements to bring consensus on ethical best practices for artificial intelligence.

However, given the imminence of enforceable legislation, being familiar with such international standards and recommendations, which often emerge from a diversity of perspectives and cultures, will provide organisations with adaptability to forthcoming regulatory requirements.

Companies should be familiar not only with the work in this space, but also with the specific standards that apply to their sectors. As industry specific standards and compliance frameworks become commonplace, a level of organisation competence to translate them into everyday processes will be needed.

3. Explainability to understandability, prospect to procedure

Attempts to provide a technical explanation of complex algorithmic behaviour has yet to produce well-placed trust in artificial intelligence systems.

The mechanisms of understandability that emphasise context and pave the way for transparency between stakeholders have similar characteristics to the reporting now used in environmental sustainability, although the disclosure of ethical harms from artificial intelligence remains further behind.

In both spaces, investor and consumer attention is focusing on potential harms. Short to medium-term business value can still be realised but, in the long term, reporting on ethical or sustainability-related impacts must be just as important as financial disclosures. A range of mechanisms can enable well-placed trust, and an understanding of what users and customers really need is crucial to deciding which are most appropriate; whether it's social context, technical explanations, transparency statements, certification or all of the above.

What this might mean for trust?

Both the cultivation of cross-organisational ethical responsibility and the inclusive approaches to standardisation, touched on in sections one and two, can help foster trusting relations among teams that develop AI. They also offer a greater overall trust between the public and artificial intelligence companies.

The practices and mechanisms described in the third section pave the way for a similar kind of trust. Yet trust from multiple stakeholders is difficult to maintain and there is a danger that these practices will provide only a layer of trustworthiness.

Perhaps the crucial transformation for the artificial intelligence ecosystem will come from a paradigm shift in which democratic governance of each individual system is embedded by design. Only then will AI systems realise their long-term business value and gain the well-placed trust of all those who shape and are shaped by them.

The reporters on this project were Luke Gbedemah
and Sam Meeson-Frizelle.