DEMOS

AI: TRUSTWORTHY BY DESIGN

HOW TO BUILD TRUST IN AI SYSTEMS, THE INSTITUTIONS THAT CREATE THEM AND THE COMMUNITIES THAT USE THEM

ELIZABETH SEGER
MARIA AXENTE

JULY 2024

This project was supported by

Published by Demos July 2024

15 Whitehall, London, SW1A 2DD
T: 020 3878 3955
hello@demos.co.uk
www.demos.co.uk

CONTENTS

ABOUT THIS PAPER

This paper is part of Demos' strategic focus area on **'Trustworthy Technology'**. With emerging technologies transforming our world at an ever faster pace, we work to build bridges between politicians, technical experts, and citizens to explore solutions, build trust, and create policy to ensure our technologies benefit society.

In this paper, written in partnership with PwC, we explore how to build trust around emerging AI technologies. AI tools promise to boost productivity, streamline business practices, and improve customer services. But these benefits will only be fully realised if the AI tools are trustworthy - if they work reliably, if they are safe, respond to consumer needs and values, and are surrounded by the reassurances of responsible institutional practices. We make the case for building systems that are trustworthy by design, not remedy. A trustworthy AI ecosystem is a common good from which all parties who yield and deal in AI technologies benefit. All AI stakeholders should appraise where in the ecosystem they sit and how they can hold up their piece of the puzzle for a future with trustworthy AI in which all parties benefit.

INTRODUCTION

There is much excitement about the opportunities of AI, to improve productivity, streamline business practices, and simplify tasks. It is also very well recognized that trust in AI will be key to fully realising these benefits; employees and public consumers will more willingly adopt and make better use of technologies they trust. Trust, however, can be a nebulous concept. People extol its virtues and study it in surveys (According to the 2024 Edelman Trust Barometer survey the AI industry is the only sector that did not experience a year-on-year boost in trust[1]) but there is a lack of clarity around what it means to have or lose trust and about how it is best achieved.

In this provocation paper we aim to demystify the concepts of trust in AI. We delineate trust from trustworthiness and emphasise the importance of putting trustworthiness first to fully realising the benefits of AI. We outline component elements of trustworthiness that work together to build an ecosystem of trust around and throughout the AI lifecycle – (1) AI tool reliability, (2) institutional processes, (3) meaningful stakeholder engagement – and we offer recommendations for how these components of trustworthiness can be pursued and demonstrated.

This paper is intentionally on the more philosophical end of the spectrum. Instead of prioritising the provision of immediately actionable steps for building trust, the primary goal is to present the reader with food for thought, to provoke contemplation on why those who develop or utilise AI technology should care about trustworthiness and on what caring about trustworthiness around AI means in practical terms.

1 Edelman, M. (March 2024). Technology's tipping point: Why now is the time to earn trust in AI.https://www.weforum.org/agenda/2024/03/technology-tipping-point-earn-trust-ai/

TRUST AND TRUSTWORTHINESS

What comes first, trust or trustworthiness? The intuitively correct answer is trustworthiness. Trustworthiness (demonstrated competence, honesty, reliability, etc.) begets trust. Trust must be earned. Yet this intuition is not well reflected in contemporary discussion. Philosopher of trust, Onora O'Neill, notes that advertisers, business and other campaigning organisations often understand trust simply as a *generic attitude* that can be observed, for instance, through polling data.[2,3] The data tells you how much an audience trusts a politician, brand, or product — this information can be useful to those looking to influence behaviour — but it does not evidence whether that trust is more or less well placed. We omit to link trust with trustworthiness, yet it is evidence of trustworthiness (of competence, honesty, reliability, etc.) that must come first as consumers look to give and refuse their trust intelligently.

Think about demonstrating trustworthiness instead of building trust.

It is important that trustworthiness be linked to trust first and foremost because misplaced trust results in harm — e.g. physical injury, financial loss, reputational damage, and broken hearts. Correspondingly, well placed trust yields benefits — e.g. the ability to comfortably delegate tasks and to benefit from the help, skills, and insights of others The aim of placing or refusing trust is therefore importantly to place more trust in those people, technologies, or institutions that are trustworthy while placing less or no trust in those who are untrustworthy.

Second, trust is slow to build yet quick to erode when disappointed. Robust trust in people, technologies, institutions and brands must, therefore, be grounded in trustworthiness. So when it comes to thinking about trust in AI as an enabler of effective AI adoption and the realisation of AI benefits, we recommend a shift in perspective. Think about demonstrating trustworthiness instead of building trust.

Finally, focussing on trustworthiness before trust has the added benefit of helping to avoid the appearance of ethics washing which sows doubt in the public eye about the genuine intentions of an organisation. Ethics washing (similar to green washing) is the phenomenon when a business exhibits seemingly good acts to gain positive brand recognition and improve product uptake and profits while those acts, in practice, have limited realised benefits. Concerns about ethics washing have been carried over to AI, with many people pointing out the quick proliferation of nice sounding ethics principles (transparency, fairness, accountability, etc.) but limited instances of effective operationalisation of those principles into practice. Many companies

2 O'Neill, O. (2020). Trust and Accountability in a Digital Age. Philosophy, 95(1), 3–17. doi:10.1017/S0031819119000457
3 O'Neill, O. (2018). Linking Trust to Trustworthiness. International Journal of Philosophical Studies, 26(2), 293–300. https://doi.org/10.1080/09672559.2018.1454637

starting with the best of intentions are blindsided by the momentous challenge of delivery; in addition to the non-ethical challenges of implementing AI tools (e.g. scaling and delivering value), translating high level principles into concrete practice requires organisational transformation that takes time and significant resources.

We will return to the intricacies of corporate governance for trustworthy AI and the role of principles in AI governance in a few pages. Presently we emphasise the importance of prioritising the establishment of ground up responsible AI practices and capabilities throughout the AI lifecycle as the necessary foundation for trust. Without this necessary action either you find yourself unprepared – when regulation comes into play, consumer expectations change, or market conditions shift, you will be playing a tough game of catch-up – or an accident that could have been forseen and prevented by more responsible practice damages your reputation.

So, think about trustworthiness before trust. To pursue trust before trustworthiness is to put the cart before the horse.

WHAT DOES TRUSTWORTHY AI LOOK LIKE?

The easy answer is, "AI tools that are worthy of trust," though it is not a particularly helpful answer. The difficulty in answering more clearly is that trustworthiness is a relative term that depends on who the trustor in question is. Who is deciding whether to give or withhold trust, why are they deciding to trust, and what are they risking in doing so?

The following chart (Table 1) works through a variety of potential trustors (or stakeholders) in AI. These stakeholders range from public consumers of services in which AI tools have been embedded, to employees that are being asked to use and integrate the new tools, to the business rolling out the new tools in their workstream, through to those companies developing and providing the AI tools. For each stakeholder, there are (a) different immediate reasons for caring about AI trustworthiness (this defines what trustworthiness means to the stakeholder), and (b) different actions that can be taken to instil confidence that the stakeholder's concerns have been addressed (to demonstrate trustworthiness according to the stakeholder's interpretation). These are not exhaustive lists but are meant to be illustrative and to invite further reflection.

Note that public perception of trustworthiness is a relevant concern for all stakeholders. Whether directly interacting with AI technologies or seeking services that employ AI tools in a workstream, the public is the ultimate consumer in the AI lifecycle and will drive demand.

TABLE 1
REASONS FOR CARING ABOUT TRUSTWORTHINESS BY STAKEHOLDER

STAKEHOLDER	(A) WHY IS TRUSTWORTHINESS AROUND AI IMPORTANT TO THIS STAKEHOLDER?	(B) WHAT IS NEEDED TO DEMONSTRATE TRUSTWORTHINESS TO THE STAKEHOLDER?
Public consumers of AI or AI-enabled services. (B2C)	• AI presents social benefits while harms e.g. from discrimination or privacy breaches are minimised. • Services are actually being improved by new AI tools. • AI-enabled services are fair and nondiscriminatory.	• Demonstrated benefit • Values reflected and respected • Needs being heard and addressed • Functional technology • Companies subject and responsive to regulation to mitigate risk
Employees using AI in their workstream (B2B)	• Impact on work experience and personal wellbeing • Worries about job security and displacement	• Worries being heard and addressed • Upskilling opportunities that take away risk of unemployment • Taking seriously AI transition as part of worker health and safety • Functional technology • Companies subject and responsive to regulation to mitigate risk
Business AI Consumers (B2B)	• Technology and service adoption by the public • Improved employee productivity • Liability concerns. Risk of adopting AI tools	• All of the above • Clear and function regulation that protects from AI associated risks risks without being overburdensome • Functional / Reliable technology
Company AI Providers (B2B or B2C)	• Adoption of AI technologies by other businesses (B2B) or Public (B2C) consumers • Liability Concerns. Risk of selling faulty tools	• All of the above
Investors	• Business growth from productive AI adoption • AI adoption bringing new opportunities	• All of the above • Demonstrated resilience to risks and liability that the sale or employment of AI tools may bring • Consistent improvements to productivity or sales (which builds off successful AI adoption)
Government	• State security (some AI applications pose threats to state security through malicious or irresponsible use) • Citizen well-being (ensure AI benefits citizens e.g. through new health application, while mitigating risks e.g. of discrimination or to data privacy)	• All of the above • Adherence by organisations to responsible AI guidelines and state regulations • Political actors adhering to best practice in the use of AI in democratic processes[4]

4 Demos (April 2024). Open Letter calling for UK political parties to safeguard election integrity in era of AI, https://demos.co.uk/research/open-letter-to-uk-political-parties-to-safeguard-the-next-general-election-from-generative-ai/

HOW DO YOU DEMONSTRATE TRUSTWORTHINESS?

We have established that trustworthiness is a multifarious concept, but this does not make the challenge of demonstrating trustworthiness around AI intractable. We recommend thinking about trustworthiness deriving from three categories: (1) AI tool reliability, (2) institutional processes, and (3) meaningful stakeholder engagement. While these categories interconnect in practice, the breakdown provides a useful starting point for discussion.

For each category, we break the process of demonstrating trustworthiness around AI into two steps:

a. **Acting:** instituting mechanisms to facilitate the production of reliable AI tools and their responsible implementation, maintenance and use. These mechanisms should include feedback mechanisms that allow those people interacting with or being impacted by the tool (developers, users, employees, service consumers) to feedback their experiences and concerns and have them addressed.

b. **Communicating:** providing users and consumers with the information they need to make informed decisions about placing or refusing their trust. Communication will in part be achieved by the above-mentioned feedback mechanisms.

We present recommendations of acting and communicating to help demonstrate trustworthiness in a table within each section (tables 2, 3 and 4). These tables are not exhaustive, but indicative of the kinds of options business might pursue.

1. AI TOOL RELIABILITY

The concepts of trust and trustworthiness are closely related to the concepts of reliance and reliability. Indeed almost all philosophers of trust agree that trustworthiness is reliability plus something extra. Though there is very little agreement on what that something extra is – goodwill, honesty, aligned values etc. But from a practical standpoint we don't need to figure it out because, as discussed, it really depends on who is doing the trusting and why.

Whether you are a producer, procurer, or general user of AI tools, the first question to ask is: How confident are you in your AI tool's outputs? The AI technology or service you are providing to consumers or for employees to use must do what it is meant to do. Is it accurate and consistent? Does it work well in a variety of contexts and across a wide user base? Does it display biases in its performance?

Reliability is established through repeated testing, evidenced through track record of past performance, and can be communicated to new users, for example, through the provision certifications - where recognised certifying bodies exist - and incident reports. The UK is building an AI Assurance ecosystem that will provide more guidance on how to validate and verify AI systems, as well as how to communicate the results of those activities.[5]

These charts that list action and are not exhaustive lists but are meant to be illustrative and to invite further reflection.

5 Introduction to AI assurance (Feb 2024). https://www.gov.uk/government/publications/introduction-to-ai-assurance/introduction-to-ai-assurance

TABLE 2
ACTING AND COMMUNICATING RELIABILITY

(A) ACTING	• Developers should conduct pre-release testing in a wide variety of contexts (e.g. as required in the EU AI Act).[6] • AI providers can engage in post-market monitoring and surveillance (e.g. as required in the EU AI Act).[7] • Deployers can roll out the new tool in stages testing its performance and reception along the way (e.g. staged release).[8]
(B) COMMUNICATING	• Certifications • Indexes • Maintain AI incident reports • Public registries of AI models

As a component of trustworthiness, AI reliability is relatively straightforward, but is only a small part of the picture. First, AI is an emerging and evolving technology that is unfamiliar to many. Compared to other technologies we use in our daily lives – from our toaster ovens to accounting software – the reliability of Generative AI tools in particular are not well-established, or where established, not long-established.[9] Many stakeholders will be hesitant to trust AI tools because they are on unfamiliar ground, with their uncertainties further fueled by a continuous stream of AI incidents reports, media frenzy, and unsubstantiated claims about AI capabilities and risks. Indeed, as public awareness of emerging AI capabilities and applications have risen, so too have public anxieties.[10]

Second, even where systems seem to perform well, system opacity poses another hurdle to adoption. When performance is imperfect (and perfection can never be guaranteed), accountability in decision-making is necessary. The responsible party must be able to provide an explanation for the decision made that is tailored to the recipient's information needs and that can be appraised if a challenge is raised. However, where AI tools are used in the delivery of professional services (e.g. processing insurance claims), given the current state of the technology, the portion of the process executed by an AI will not be transparent to the user, and "because the AI said so" is not a sufficient response for a decision made.

6 https://artificialintelligenceact.eu/chapter/3/
7 https://artificialintelligenceact.eu/chapter/9/
8 Solaiman et al. (2019). Release Strategies and the Social Impacts of Language Models. https://arxiv.org/abs/1908.09203
9 AI encapsulates a broad range of technologies ranging from narrow machine learning (ML) applications applied in restricted contexts (e.g. ranking algorithms, content recommendation systems, diagnostic system) to much broader general AI systems, also called foundation models, that can function across a wide variety of tasks, and application spaces. Generative AI systems which produce original content (images, video, audio, and or text) are the most recent innovation in foundation model development with promising applications across sectors. On the spectrum, from narrow ML applications to generative AI, narrow MLs have more well established use cases in industry, clearly documented track records within specific use contexts, and employ straightforward algorithmic processes to derive solutions. Narrow ML application, like most tools, can be harmful if applied outside their intended contexts or where training biases are overlooked or ignored. On the other end of the spectrum, greater excitement for the transformational potential of AI sits with foundation models and generative AI, but their most promising use cases are all less-well established, the technology and its capabilities and risks are continuously evolving, and the processes leading from system input to output are more opaque.
10 Public attitudes to data and AI: Tracker survey (Wave 3). https://www.gov.uk/government/publications/public-attitudes-to-data-and-ai-tracker-survey-wave-3/public-attitudes-to-data-and-ai-tracker-survey-wave-3

SHIFTING TO AN "AI IN THE LOOP" MINDSET

You have probably heard about "human in the loop" as a strategy for maintaining accountability. The idea is that for any autonomous decision making process, a human actor is inserted into the picture as the responsible party. Their job is to confirm outputs and sign off on key decision points. The challenge with the human-in-the-loop model, however, is that it is not clear that a human can be inserted in any meaningful way. If it is the case that AI systems have an opaque decision-making process that humans would struggle to understand even if laid bare, then how could a human be held accountable as the checks on those decisions? Indeed, what human would want the responsibility of being the human in the loop? Furthermore, if the purpose of implementing the AI tool in the first place is to improve on human performance, why would we then wish to fall back on human decisions?

At our current stage of AI development where new AI capabilities are emerging and evolving and both users and downstream service consumers are still digesting the idea of AI implementation to potentially high-stakes contexts, we suggest shifting to an "AI in the loop" perspective. "AI in the loop" flips the concept on its head. Instead of inserting a responsible human into a primarily AI run process, the idea is that we insert AI tools into human processes to fulfil tasks where the AI is clearly better suited and build it into our workflows. This will primarily be to help with laborious tasks (e.g. finding and summarising information), but not judgement tasks. The process remains human led, rooted in human accountability, and aligned with human needs and values from start to finish with AI tools strategically implemented to improve human performance, not the other way around.

AI EXPLAINABILITY AND TRUST
PWC PARTNER INSIGHT*

Explainability continues to be a challenge across different sectors. When organisations use an Excel spreadsheet or any other IT system, the outputs are deemed acceptable because they offer transparent insight into how the answers were derived. However, when using AI, especially generative AI, the process of how the answers are generated is often not understood, leading to considerable nervousness about using AI for supporting sensitive decision making. There are specific use cases where the decision-making process of an AI system is more transparent. These cases involve controlling the input (e.g., providing the system with a set of PDFs) and defining the question (e.g., summarising key points related to X). Practitioners can leverage their expertise to evaluate the quality and realism of the output based on the provided inputs and prompts.

In the Technology, Media, and Telecommunications (TMT) sector, large teams draft extensive bespoke contracts. Generative AI is well-suited for reading, summarising, and processing documentation, presenting a significant opportunity to improve productivity. However, legal teams and clients express concerns about the lack of transparency in AI decision-making, hindering the assessment of AI's comprehensiveness in considering risks and opportunities. To address liability and build trust, an 'AI in the loop' approach is recommended. This involves utilising AI for discrete tasks where it excels, such as summarisation, while reserving judgement-based tasks for human practitioners

* PwC insight boxes provide a view into PwC experiences working with clients using and adopting AI tools.

Finally, trustworthiness is not just about how the AI tool performs but about how it is being integrated into services and, given that perfect performance is never guaranteed, how risks are being managed and mitigated. The "something extra" of trustworthiness over reliability rests with the people and processes surrounding the technology. AI is a sociotechnical system; which AI tools people choose to use and how they use those tools will have an impact on trust.

2. INSTITUTIONAL PROCESSES

Absence of guaranteed performance and unfamiliarity with decision making-processes are not unique challenges when deciding whether to trust external entities. We navigate the minefield every day with our doctors, mechanics, lawyers, and architects. When you go to see a new GP, for example, you have no reason to trust that doctor based on what you know about them as an individual. You've never met them before and have no direct knowledge of their track record for successful diagnosis or treatment. More so, given your limited medical knowledge, you are not well positioned to make a meaningful appraisal of any explanation the physician would give for a diagnosis or treatment plan - it wouldn't give you a strong justification for your belief or disbelief in the GP's claims (this difficulty of explanation exists in any expert-novice relationship. The greater the gap in expertise between an expert and the person seeking the expert's advice, the weaker the epistemic justification for belief the novice can derive from the explanation).[11] Nonetheless you deem the GP trustworthy enough as a medical practitioner to task them with your care.

You base this trust on the GP's membership to the medical profession and employment in a medical office. They must, you assume, have completed rigorous training, maintain up-to-date knowledge of medical advances, meet sufficiently high standards of demonstrated competence, and, as is standard in the medical profession, be committed to principles of nonmaleficence and beneficence. Your trust in the GP is overwhelmingly grounded in the perceived trustworthiness of the institutional framework in which they are embedded. If you had reason to believe those institutional supports and controls were slipping – that the institutional trustworthiness was slipping – then so too would your trust in any first-visit GP.

Trust in AI is much the same. With limited knowledge of the AI tools being used, we must rely on our assumptions that along the chain of AI development, deployment, and use in the provision of services, that institutional mechanisms are in place to ensure safety and responsibility at each stage. Trustworthy AI is the product of a trustworthy ecosystem. However, much nervousness about the adoption of AI, especially into sensitive industries like healthcare and finance, stems precisely from the worry that such systemic controls are weak or do not exist.

Voluntary frameworks and nascent state backed regulation have yet to deliver towards consumer trust, and so all private and public stakeholders along the AI value chain have a serious responsibility to maintain that institutional trustworthiness. They also have strong interest in doing so to help inform policy that mitigates risks to protect themselves and their consumers without being overly burdensome.

Trustworthy systems must be trustworthy by design, and this means building in responsibility from the top-down, bottom-up, and throughout the AI lifecycle.

With respect to trustworthy institutional processes, how confident are you as a private or public entity in the processes your organisation has in place as contributing to a trustworthy ecosystem of intervention and control around AI? The key is embedding best practices throughout AI development and deployment.[12] Systemic trustworthiness requires more than vetting data sets for biases, or running post mortem audits. It's a matter of moving these processes to the very start of your operation and following them through at every step so that as soon as you start working with AI, you are following best practice. This way red flags emerge as you go, to point you in the right direction. It's a significant investment, but as an operation scales, the damage of catching problems late will also scale, so the benefits pay off.

11 Seger, E. (2022). Ch. 2 Continuums of Justificatory Value. In E. Seger (Ed.), Experts & AI systems, explanation & trust: A comparative investigation into the formation of epistemically justified belief in expert testimony and in the outputs of AI-enabled expert systems (p. 57-73). https://doi.org/10.17863/CAM.90175

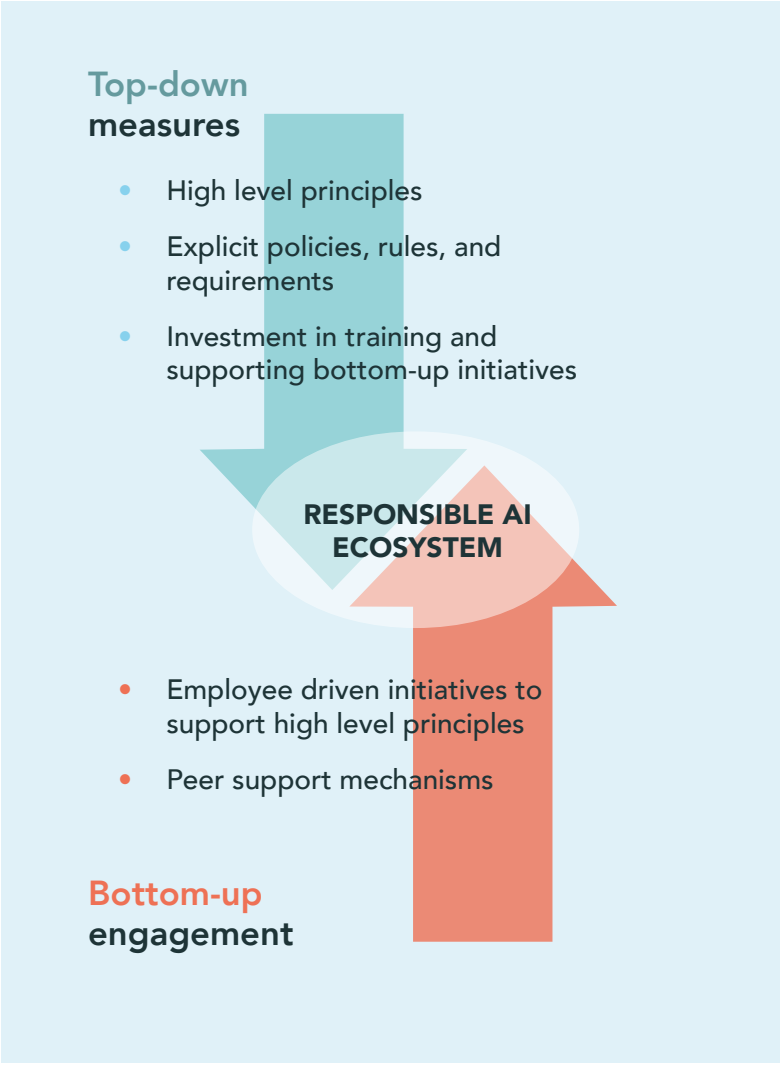
12 PwC (2024). From principles to practice: Responsible AI in action. https://www.strategy-business.com/article/From-principles-to-practice-Responsible-AI-in-action

Trustworthy systems must be trustworthy by design, and this means building in responsibility from the top down, bottom up, and throughout the AI lifecycle.[13] From the top down, organisation leadership sets the tone, providing guidance in the form of high-level principles (e.g. transparency, explainability, accountability etc.) that are reinforced through the provision of training, and explicit rules and requirements that operationalise those principles. From the bottom-up, employees are involved in the construction and refinement of those explicit guidelines, rules and requirements. Such bottom-up efforts include employee-led initiatives to identify opportunities for operationalising principles, peer-to-peer support programs to aid in adherence, and community involvement in the articulation of rules and requirements to ensure they are fit for purpose and not unnecessarily burdensome.

Without bottom-up expert involvement, top-down efforts to introduce and reinforce high-level principles can struggle to find traction. Indeed, where top-down rules and regulation are felt to be introducing superfluous busywork, resentment can fester, which may in turn lead to community wide demoralisation and resistance to the very principles that are meant to be reinforced.[14] So involving bottom-up employee involvement in development, administration, and continued review of responsible AI measures is key to their effective update and organisation-wide ownership.

Meanwhile, top-down involvement from organisation leadership is needed to provide the resources and support necessary for bottom-up initiatives. Top-down support allows utilisation of established organisational infrastructure and monetary resources and offers validation for the values and agendas being pushed. Without it, bottom-up initiatives will struggle to impact responsible AI development and use and are more likely to lose momentum.

FIGURE 1
BUILDING IN RESPONSIBILITY FROM THE TOP-DOWN AND BOTTOM-UP



Communicating trustworthy institutional processes

To complete the process of *demonstrating* trustworthiness, the implementation of these top-down and bottom-up measures must be communicated throughout the stakeholder chain, within the organisation, to investors and the public. This can be achieved through active feedback mechanisms between stakeholders, through transparent record keeping, and public engagement initiatives.

13 Seger, E. (2022). Ch. 5 Well-functioning systems. In E. Seger (Ed.), Experts & AI systems, explanation & trust: A comparative investigation into the formation of epistemically justified belief in expert testimony and in the outputs of AI-enabled expert systems (p. 106-136). https://doi.org/10.17863/CAM.90175

14 Pettit, P. (2002). Instituting a research ethic: Chilling and cautionary tales. In P. Pettit (Ed.), Rules, Reasons, and Norms. Oxford Scholarship Online: Oxford University Press.

TABLE 3
ACTING AND COMMUNICATING TRUSTWORTHY INSTITUTIONAL PROCESSES

(A) ACTING	• Top down measures (e.g. model testing processes, training requirements, high-level principles, risk assessment guidelines) • Bottom up measures (e.g. peer support programs, employee led governance initiatives)
(B) COMMUNICATING	• Transparency • Active Feedback mechanisms • Record keeping

A note on high-level principles

As noted earlier in the paper, a significant investment of time and resources is needed to operationalise high-level AI principles. As such, there is a worry that the risk of unwarranted visual signalling or "ethics washing" is high among those organisations that claim to adopt principles to guide their practice.[15] While there likely exist some organisations that do post principles on their walls and leave their responsible AI efforts at that, for the vast majority it is more often that the intentions are genuine, but translating principles to practice is a gargantuan challenge. As originally outlined by one of this paper's authors, Elizabeth Seger, one way to approach this challenge is to think of principles as having two clear functions: a start-point function and cultural influence function.[16]

1. Start Point Function:

The first and most straightforward function of high-level principles is to serve as a start point for articulating more explicit rules and regulations to direct responsible practice throughout the AI lifecycle. By virtue of their broad nature, principles do not offer specific, ground-level guidance on their own. Nonetheless, they serve a valuable purpose in categorising ethical considerations for further investigation. They provide a common point of departure for deliberating and articulating more explicit rules and processes that should be put in place to sustain responsible AI development, deployment, and use.

Of course progressing from the articulation of high-level principles to the articulation and implementation of specific practices still takes significant investment of time and resources. However worries about this hurdle that lead some to dismiss the value of principles and cite concerns about ethics washing tend to overlook a second key function of high-level principles.

2. Cultural Influence Function:

An often overlooked yet equally crucial role of AI ethics principles is shaping and influencing cultural norms and values. Principles provide a common, guiding vocabulary with which AI developers and employers discuss the challenges they face and contemplate potential impacts, risks, and opportunities. Where new principles challenge the status quo, they can help catalyse cultural shift. For example, the fast-paced Silicon Valley ethos primarily extols the virtues of efficiency, optimization, and scale while many proposed principles, such as fairness, accountability, explainability, inclusivity, and transparency, challenge this prevailing mindset and nudge culture towards prioritising responsibility and human welfare. Viewed as a tool for framing mindsets and nudging cultural change, it does not so much matter how, exactly, those principles are defined, but rather that they are consistently engaged with and widely discussed and debated.

Why care about culture? If explicit rules and requirements are the letter of the law, then culture is the spirit. Culture fills in the gaps where explicit rules and regulation fall short, helping practitioners to make decisions that align with organisational goals and values on their own accord. More so, cultural alignment improves the uptake and efficacy of explicit guidelines. Individuals are more driven to fully adhere to organisation policies and further organisation goals if those policies and goals resonate with the cultural norms and values already internalised by the communities to which they are being applied.

15 Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1, 501–507. https://doi.org/10.1038/s42256-019-0114-4

16 Seger, E. (2022). In defence of principlism in AI ethics and governance. Philosophy & Technology, 35(2), 45. https://doi.org/10.1007/s13347-022-00538-y

ORGANISATION CULTURE AND REBUILDING TRUST LOST
PWC PARTNER INSIGHT*

Fostering a culture of responsibility and excitement among employees is key to operationalising organisational goals whether those goals are to build more responsible AI systems, to stretch the frontier of innovation, or to build more inclusive work environments. When corporate culture aligns with corporate goals you will experience smoother uptake of changes to rules and requests in pursuit of those goals, and employees will be more driven to meet if not exceed expectations.

When consumer trust is damaged by an accident or an unforeseen event, businesses must prioritise transparency as the essential first step. Accountability and transparency are crucial in restoring trust. In the context of AI, transparency encompasses both technical aspects, such as the ability to explain the decisions made by a model, and organisational practices, including how transparent institutions are about their AI processes and decision-making. To rebuild trust there are three steps with institutional transparency at their core:

1. **Own the problem.** Publicly acknowledge fault where it sits.
2. **Diagnose the problem and formulate a plan** to address it. Be transparent about that plan.
3. Most importantly, **hold to that plan.** Show consumers that you have taken the issue seriously and have executed the steps you identified.

Sandra Sucher and Shalene Gupta's book, *The Power of Trust: How Companies Build It, Lose It, and Regain It Again* (2021), is essential reading on the subject.

3. Meaningful stakeholder engagement

So far in this paper we have discussed two sources from which AI trustworthiness derives. The first is AI reliability (does the AI tool you've produced or employ work?) and the second is institutional process (do the organisations that develop or employ AI tools have robust responsible AI practice and culture running throughout the AI lifecycle?). This section turns to a third factor - the role of stakeholder engagement.

When you go to a GP, your trust in GP is only partly based on your belief that they are well trained and held to standards enforced by their profession. In another large part, your trust is based on the assumption that they have your best interests at heart. This is why, when a GP rushes through the niceties at the start of an apportionment or brushes aside your questions, your trust likely starts to dwindle. Your faith is being shaken, not necessarily in their medical expertise or skill, but in their care for you as an individual. Do they understand your priorities, needs, and worries? And if not, will their advice be the best advice for your specific situation?

Imagine, for example, that you need surgery on your ankle. Your doctor recommends a procedure that is long lasting and has a quick recovery time. It will limit your ankle mobility a little, but nothing you should notice on the day to day, and all the pain will be resolved. Most people would opt for this option, and so that is what your physician recommends. There is, however, an alternative procedure that guarantees full ankle mobility, but will need to be repeated every five years and will result in persistent pain. You are a professional ballet dancer, and your art (impossible with any limit on ankle mobility) is everything to you. You would have opted for the second option had the option been given. In this case your doctor

* PwC insight boxes provide a view into PwC experiences working with clients using and adopting AI tools.

disappointed your trust, not because of any failing of medical expertise, but because of a failure to engage with your needs and values and advise accordingly.

Stakeholder engagement and responsiveness values and concerns is key to trustworthiness. This holds whether you are a physician or a provider of AI tools or AI-enabled services.

It is about ensuring that AI is something done for and with people, not something done to them.

At the end of the day, AI tools are meant to impact human lives – to improve productivity, transform public service, revolutionise transportation, solve complex scientific problems, and provide information and entertainment. So it is essential that the needs and values of those people being impacted by the technology are being taken into account. It is about ensuring that AI is something done for and with people, not something done to them. This concept of Human Centred AI (HCAI) is one coined by the respected AI scholar Shannon Vallor to describe AI developed and deployed with people, for people, and by people.[17] The HCAI concept is currently missing from industry, but we posit it is a missing link that can bridge short term organisational objectives for using AI to improve productivity and efficiency with the longer term and broader ambitions of building and sustaining trust in AI and AI-enabled services.

Human Centred AI is ultimately pursued by engaging downstream stakeholders (users and consumers) in decision-making about the development, deployment, and employment of AI tools. This engagement plays a dual function both in helping to build more trustworthy AI tools and AI-enabled services that better serve consumer needs and value, and in communicating that trustworthiness to the consumers. In this way, meaningful stakeholder engagement is good business. It's morally good business - it puts people at the heart of a technology that will have profound impacts on their lives. And it's good *for* business – well-grounded trust facilitates more willing and effective technology adoption and more widespread realised benefits from the technology.

In a review of evidence feeding into the UK AI summit, the Ada Lovelace institute found that "Taking into account people's perspectives and experiences in relation to AI – alongside expertise from policymakers and technology developers and deployers – is vital to ensure AI is aligned with societal values and needs, in ways that are legitimate, trustworthy and accountable." They continue, "public views point towards ways to harness the benefits and address the challenges of AI technologies, as well as to the desire for diverse groups in society to be involved in how decisions are made."[18]

Of course there are some practical challenges to involving consumers in decision-making about AI. It's not possible, for instance, to sit a representative sample of the public at the elbow of an AI developer to guide each step as they code. Some companies will also have IP concerns about sharing proprietary information about their products with a wider audience of business or public consumers.

That said, there are ways to meaningfully involve stakeholders in principle setting and decision making about AI that do not require involvement in the development process itself (e.g. by consulting stakeholders in prioritising opportunities for application development, and defining contexts of appropriate use). The table below is by no means an exhaustive list of methods for stakeholder engagement, but is a starting point to build on. Importantly, these methods will only be effective if accompanied by commitments from decision-makers to embed the engagement processes in their decision-making procedures.

17 Vallor, S. (2024). Defining Human-Centered AI. In Regis, C., Denis, J., Axente, M., and A. Kishimoto (Eds.), Human-Centered AI: A Multidisciplinary Perspective for Policy-Makers, Auditors, and Users. https://www.routledge.com/Human-Centered-AI-A-Multidisciplinary-Perspective-for-Policy-Makers-Auditors-and-Users/Regis-Denis-Axente-Kishimoto/p/book/9781032341613#

18 Reeve, O., Colom, A., & Modhvadia, R. (2023). What do the public think about AI? https://www.adalovelaceinstitute.org/evidence-review/what-do-the-public-think-about-ai/#finding-6-there-are-important-gaps-in-research-with-underrepresented-groups-those-impacted-by-specific-ai-uses-and-in-research-from-different-countries-24

TABLE 4
ACTING AND COMMUNICATING MEANINGFUL STAKEHOLDER ENGAGEMENT

(A) ACTING	• Consumer consultation to inform initial product development (What AI applications are being developed and how are they serving consumer needs) • Stakeholder consultation on acceptable risks thresholds • Product/service feedback mechanisms • Product/service support • Corporate governance structure (e.g. having a democratically elected board or incorporating as a public benefits company)
(B) COMMUNICATING	• The above play a dual function of facilitating stakeholder engagement and also serving as means of communication

ETHICAL AI PAVES THE WAY FOR GRADUAL AND RESPONSIBLE AI ADOPTION

PWC PARTNER INSIGHT*

Our experiences working in Financial Services (FS) have demonstrated just how important applied ethics is in the context of a responsible AI framework. Just because you can build an AI application or employ an AI tool doesn't not mean you should. The first step for any business considering developing or adopting AI technologies is to be very clear about why you are building or adopting the technology. What are your intentions? What do you want to achieve? How would you measure success (e.g. in improved productivity, customer satisfaction, etc.)? And what risks and ethical dilemmas could you expect from the tools application? The aim is to start working in mitigations to those risks from the very beginning, from the ideation phase onward. This will result in better AI tools and implementations and minimise instances of shocks to the system from problems that emerge down the line.

Being very clear about the goals for implementing AI and its impact and implications is also key to building trust among stakeholders. AI must provide clear benefits beyond the status quo. Do not adopt AI solely for its novelty. A gradual and cautious approach to AI integration is essential for fostering trust. This allows for the thorough evaluation of its benefits and challenges, while also incorporating feedback from employees and consumers

This kind of gradual adoption has been working well in the financial sector, which has had the maturity and appetite to adopt responsible AI governance frameworks over the years. The financial sector has commenced extensive proof-of-concept studies to assess impacts, gather feedback, and make improvements. It is now beginning to scale AI tools for broader and more general applications.

* PwC insight boxes provide a view into PwC experiences working with clients using and adopting AI tools.

One of the most successful AI applications in the banking sector is the management of client feedback. AI gathers customer comments and complaints from various channels, such as transcribed phone calls, online chats, and in-branch communications. It then identifies emerging themes and the root causes of complaints. The AI drafts suggested action plans for addressing issues, which can be communicated back to customers. Whereas responding to complaints with actionable plans previously took several weeks, tailored responses are now issued within 24 hours. These AI tools have enhanced the ability of banks to meet diverse customer needs and significantly improved the customer experience.

Concerns often arise that responsible AI governance may hinder innovation. However, a cautious approach is essential for building trust. Currently, there is significant progress in making AI governance more dynamic. This involves fostering strong cultures of responsible AI within organisations, providing training to raise awareness of potential challenges and risks, and clearly defining the goals of new AI tools from the outset to ensure risk resilience.

CONCLUSION

This provocation paper has aimed to demystify the concept of trust: to delineate between trust and trustworthiness, to emphasise the importance of putting trustworthiness first before trust, and to illustrate how the different elements of reliability, institutional process, and stakeholder engagement work together to build trustworthiness throughout the AI lifecycle. It makes the case for building systems that are trustworthy by design, not remedy.

What makes you confident in the reliability of the AI tools you produce or use? Do the institutional processes you have in place support responsible practice throughout the AI lifecycle and company culture? What methods do you employ to engage and implement feedback from external stakeholders?

The trustworthy AI ecosystem is a common good from which all AI stakeholders benefit.

The trustworthy AI ecosystem is a common good from which all AI stakeholders benefit, and when public confidence in the reliability and social beneficence of AI technologies falters, all parties who produce or employ AI tools in the provisions of services will feel the effects and business and well-being. We encourage readers to contemplate where in the AI lifecycle they sit as a stakeholder (Table 1) and how they are pursuing each of the elements of trustworthiness to hold up their piece of the puzzle.

Licence to publish

Demos – Licence to Publish

The work (as defined below) is provided under the terms of this licence ('licence'). The work is protected by copyright and/or other applicable law. Any use of the work other than as authorized under this licence is prohibited. By exercising any rights to the work provided here, you accept and agree to be bound by the terms of this licence. Demos grants you the rights contained here in consideration of your acceptance of such terms and conditions.

1 Definitions

a 'Collective Work' means a work, such as a periodical issue, anthology or encyclopedia, in which the Work in its entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole. A work that constitutes a Collective Work will not be considered a Derivative Work (as defined below) for the purposes of this Licence.

b 'Derivative Work' means a work based upon the Work or upon the Work and other pre-existing works, such as a musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Work may be recast, transformed, or adapted, except that a work that constitutes a Collective Work or a translation from English into another language will not be considered a Derivative Work for the purpose of this Licence.

c 'Licensor' means the individual or entity that offers the Work under the terms of this Licence.

d 'Original Author' means the individual or entity who created the Work.

e 'Work' means the copyrightable work of authorship offered under the terms of this Licence.

f 'You' means an individual or entity exercising rights under this Licence who has not previously violated the terms of this Licence with respect to the Work, or who has received express permission from Demos to exercise rights under this Licence despite a previous violation.

2 Fair Use Rights

Nothing in this licence is intended to reduce, limit, or restrict any rights arising from fair use, first sale or other limitations on the exclusive rights of the copyright owner under copyright law or other applicable laws.

3 Licence Grant

Subject to the terms and conditions of this Licence, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) licence to exercise the rights in the Work as stated below:

a to reproduce the Work, to incorporate the Work into one or more Collective Works, and to reproduce the Work as incorporated in the Collective Works;

b to distribute copies or phono-records of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission the Work including as incorporated in Collective Works; The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. All rights not expressly granted by Licensor are hereby reserved.

4 Restrictions

The licence granted in Section 3 above is expressly made subject to and limited by the following restrictions:

a You may distribute, publicly display, publicly perform, or publicly digitally perform the Work only under the terms of this Licence, and You must include a copy of, or the Uniform Resource Identifier for, this Licence with every copy or phono-record of the Work You distribute, publicly display, publicly perform, or publicly digitally perform. You may not offer or impose any terms on the Work that alter or restrict the terms of this Licence or the recipients' exercise of the rights granted hereunder. You may not sublicence the Work. You must keep intact all notices that refer to this Licence and to the disclaimer of warranties. You may not distribute, publicly display, publicly perform, or publicly digitally perform the Work with any technological measures that control access or use of the Work in a manner inconsistent with the terms of this Licence Agreement. The above applies to the Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Work itself to be made subject to the terms of this Licence. If You create a Collective Work, upon notice from any Licensor You must, to the extent practicable, remove from the Collective Work any reference to such Licensor or the Original Author, as requested.

b You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended

for or directed toward commercial advantage or private monetary compensation. The exchange of the Work for other copyrighted works by means of digital file sharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in connection with the exchange of copyrighted works.

c If you distribute, publicly display, publicly perform, or publicly digitally perform the Work or any Collective Works, you must keep intact all copyright notices for the Work and give the Original Author credit reasonable to the medium or means You are utilizing by conveying the name (or pseudonym if applicable) of the Original Author if supplied; the title of the Work if supplied. Such credit may be implemented in any reasonable manner; provided, however, that in the case of a Collective Work, at a minimum such credit will appear where any other comparable authorship credit appears and in a manner at least as prominent as such other comparable authorship credit.

5 Representations, Warranties and Disclaimer

a By offering the Work for public release under this Licence, Licensor represents and warrants that, to the best of Licensor's knowledge after reasonable inquiry:

i Licensor has secured all rights in the Work necessary to grant the licence rights hereunder and to permit the lawful exercise of the rights granted hereunder without You having any obligation to pay any royalties, compulsory licence fees, residuals or any other payments;

ii The Work does not infringe the copyright, trademark, publicity rights, common law rights or any other right of any third party or constitute defamation, invasion of privacy or other tortious injury to any third party.

b Except as expressly stated in this licence or otherwise agreed in writing or required by applicable law, the work is licenced on an 'as is' basis, without warranties of any kind, either express or implied including, without limitation, any warranties regarding the contents or accuracy of the work.

6 Limitation on Liability

Except to the extent required by applicable law, and except for damages arising from liability to a third party resulting from breach of the warranties in section 5, in no event will licensor be liable to you on any legal theory for any special, incidental, consequential, punitive or exemplary damages arising out of this licence or the use of the work, even if licensor has been advised of the possibility of such damages.

7 Termination

a This Licence and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this Licence. Individuals or entities who have received Collective Works from You under this Licence, however, will not have their licences terminated provided such individuals or entities remain in full compliance with those licences. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this Licence.

b Subject to the above terms and conditions, the licence granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different licence terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this Licence (or any other licence that has been, or is required to be, granted under the terms of this Licence), and this Licence will continue in full force and effect unless terminated as stated above.

8 Miscellaneous

a Each time You distribute or publicly digitally perform the Work or a Collective Work, Demos offers to the recipient a licence to the Work on the same terms and conditions as the licence granted to You under this Licence.

b If any provision of this Licence is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this Licence, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

c No term or provision of this Licence shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

d This Licence constitutes the entire agreement between the parties with respect to the Work licenced here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This Licence may not be modified without the mutual written agreement of Demos and You.

DEMOS

Demos is a champion of people, ideas and democracy. We bring people together. We bridge divides. We listen and we understand. We are practical about the problems we face, but endlessly optimistic and ambitious about our capacity, together, to overcome them.

At a crossroads in Britain's history, we need ideas for renewal, reconnection and the restoration of hope. Challenges from populism to climate change remain unsolved, and a technological revolution dawns, but the centre of politics has been intellectually paralysed. Demos will change that. We can counter the impossible promises of the political extremes, and challenge despair – by bringing to life an aspirational narrative about the future of Britain that is rooted in the hopes and ambitions of people from across our country.

Demos is an independent, educational charity, registered in England and Wales. (Charity Registration no. 1042046)

Find out more at **www.demos.co.uk**

DEMOS

PUBLISHED BY DEMOS JULY 2024

15 WHITEHALL, LONDON, SW1A 2DD
T: 020 3878 3955
HELLO@DEMOS.CO.UK
WWW.DEMOS.CO.UK