AI Assurance Processes

Emre Kazim, Adriano Koshiyama
Department of Computer Science, University College London, UK
Contact: e.kazim@ucl.ac.uk a.koshiyama@cs.ucl.ac.uk

Abstract

The ethics of Artificial Intelligence (AI Ethics) can be thought of as undergoing three broad phases, with the first two being principles and processes, to the current phase which we read in terms of assurance. In this article we anticipate and expand upon themes in the AI assurance space: namely; i. general and sector specific assurance, ii. governance via technical and impact assessments, iii. the use of 'traffic-light' user friendly monitoring interfaces, iv. risk assessments and 'red teaming' in order to mitigate unknown risks, v. certification, and vi. insurance of AI systems. The article concludes by noting that the manner in which AI assurance matures is likely to be as a result of the extent and reach of any regulatory intervention.

Keywords: Artificial Intelligence, AI Ethics, Assurance, Audit, Impact Assessment, Certification, Red Teaming, Insurance, Regulation, Standards.

1. Introduction

The ethics of Artificial Intelligence (AI Ethics) can be thought of as undergoing three broad phases, with the first two being principles and processes, to the current phase which we read in terms of assurance. In this article we anticipate and expand upon themes in the AI assurance space: namely;

i. **General and sector-specific assurance:** here we contrast broad national regulation and standards with sector specific ones, and note that sector specific regulation and standards will involve the emphasizing/expansion of parts of the broader regulation rather than the introduction of a different set of standards.
ii. **Governance:** Here we discuss two aspects of AI governance, namely technical assessments and impact (risk, compliance, etc.) assessments.
iii. **Monitoring Interfaces:** Here we outline risk assessments and the use of 'traffic-light' (colour-code) user friendly monitoring interfaces.
iv. **Unknown Risks:** here we discuss risk schemes and highlight 'red teaming', which is used to mitigate unknown risks.
v. **Certification:** here we outline the numerous ways in which certification may occur, such as certification of a system or AI engineers, etc.
vi. **Insurance**: Here we briefly note that we anticipate insurance of AI systems to emerge as a result of the AI assurance space maturing.

We close with a conclusion noting that the manner in which AI assurance matures is likely to be as a result of the extent and reach of any regulatory intervention.

2. AI Principles to AI Assurance

With the proliferation of AI research, development and deployment, along with high-profile cases of harm, awareness of the social impact and ethical implications of AI increased within the various stakeholders – namely, the academy, government, civil society (through NGOs) and industry. Examples of harm that were observed are bias in systems such as recruitment [1] and criminal

justice sentencing (where particular demographics are prejudiced against) [2], voter manipulation [3] and misdiagnosis of cancer patients [4]. With these a growing consciousness developed within wider society and developers of these technologies that something needs to be done. Indeed, what is now referred to as 'AI ethics' or 'trustworthy AI' or 'responsible AI' is the body of literature that has resulted because of this consciousness and debate [5].

In our reading, the field of AI ethics has undergone three broad phases:

Principles: The first was an AI ethics set of principles. This resulted in a large number of academic groups, governmental bodies (both national and international), private industry and NGOs, publishing their own set of principles. The core problem and criticism of this was that these principles were largely incoherent and contradictory. Ethical consensus and clarity was elusive [6].

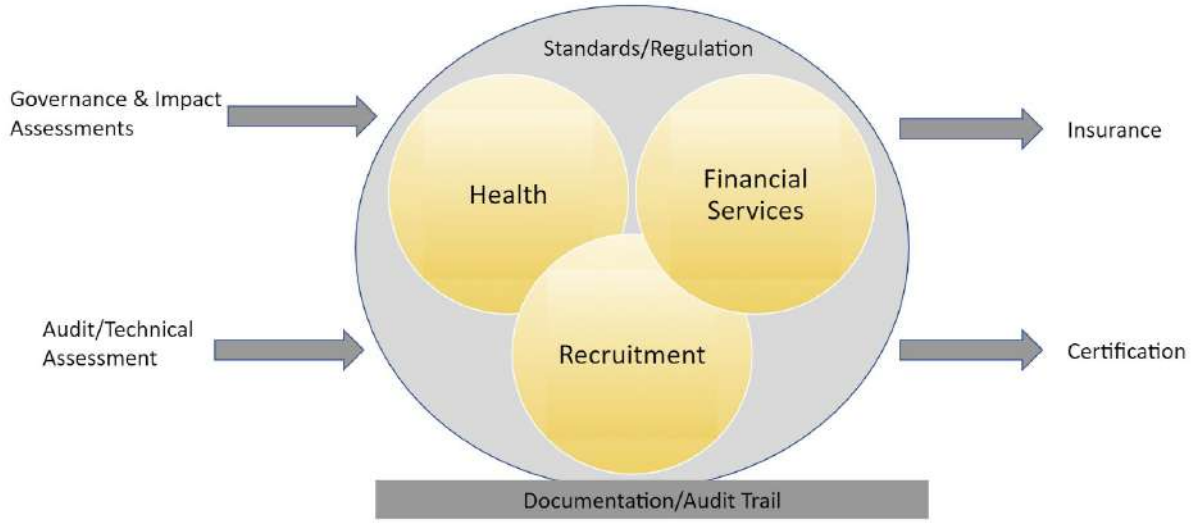
Processes: The second phase was an ethical-by-design approach, which was an engineering focused problem-solving exercise. Here the issue was two-fold: attempting to establish what it means to be ethical-by-design, and, probing the question as to whether or not engineers are best placed to both understand and judge (let alone implement) ethics. As was the case with the principles phase, no clear and consensual ethical code emerged that developers could somehow engineer into design, and the question remained unclear regarding whether the problem is indeed an engineering one [7].

Motivating this debate was the social scientific literature, where the social and ethical implications of technology are often discussed in terms of 'value laden technology' theories [8]. This concerns the set of social values that technologies sit within and express. For example, nuclear technology is used both as a source of energy and as a devastating and indiscriminate weapon. The use of the technology – in this case AI technologies - is seen as an expression of the social and political values of the society. Hence the argument can be made that ethics is a social and political concern, not an engineering one.

Finally, within this phase, and recognising that the 'problem of AI ethics' sits in an interdisciplinary overlap, the idea that only an interdisciplinary approach can meaningfully address and hope to mitigate the ethical risks of AI came to fruition. The idea here is that philosophers, sociologists, psychologists, lawyers, engineers etc. work together in a dynamic and integrated manner. Hence the governance, social impact, legal compliance, engineering standards etc. are all continuously kept in mind from the design phase to real life use.

Assurance: we read the third – indeed the current phase – as concerned with the need to standardise and operationalise the AI ethics discipline. Where we read assurance as a broad term to encompass certification and audit. The quest to achieve trustworthy AI has matured to the point where appropriate governance, regulation, impact assessment and auditability standards are being proposed and formulated.

Within the context of this third phase of AI ethics, in the following sections we anticipate and expand upon aspects of AI assurance (Schematic 1).



Schematic 1: the various aspects and order of AI assurance that we anticipate will develop.

3. General and Sector-Specific Assurance

We believe that the satisfaction of a particular standard - ex. certification, auditability, etc. - will become mandatory. We read this from the growing calls for AI systems to be appropriately governed i.e. ensuring AI systems are trustworthy. Indeed we read this from the signalling of both international [9, 10] and national regulatory [11, 12] and standard bodies [13, 14]. In this respect we anticipate that standards will be both general and specific.

General Standards: With respect to the general, we believe that guidance (which may or may not be legally codified) will encompass broad dimensions such as privacy, explainability, safety and fairness, and these will be set by institutions and bodies with non-sector specific remits (ex. the UK's Information Commissioner's Office). Developments in this space are becoming more concrete. For instance, in the recent 'Explaining decisions made with AI' , the Information Commissioner's Office & The Alan Turing Institute [15] advise on how organisations can explain the processes, services and decisions delivered or assisted by AI to those that are affected by such decisions. The guidance outlines explanations in terms of who is responsible, data choices and management, fairness considerations, safety and impact [16].

Sector Standards: Currently sector specific guidance already exists. These address idiosyncrasies of application. For example, the UK's Financial Conduct Authority is leading in the standards of AI systems in financial services [17] and the Care Quality Commission is doing so in health and social care provision. In addition to sector specific regulators issuing guidance, sectors themselves are developing their own standards and approaches to best practice. Recruitment and healthcare are examples of this [18].

4. Governance: Impact Assessment and Technical Audit

Within the context of AI systems governance can be divided into two broad streams, namely technical and non-technical.

Non-technical governance: concerns systems and processes that focus on allocating decision makers, providing appropriate training and education, keeping the human-in-the-loop, and conducting social and environmental impact assessments, all of which fall under mitigation strategies. Governance norms and sector specific particularities in the context of AI development

and product deployment will need to develop, along with the issue of accountability. Here what is being referred to is:

i. who will be liable if something goes wrong (processor, controller, user) i.e. the allocation of responsibility;

ii. what current legislation like GDPR, financial regulations, etc. have to say on a case-by-case basis; and

iii. differences between countries and economic blocks.

Within this context there is also the AI auditing literature, which calls also for doing a Data Protection Impact assessment when algorithms are used [19]. Additionally, there are calls for AI impact assessments that address issues of i. human-rights, ii. social concerns and iii. environmental [20, 5].

Technical governance: this concerns systems and processes that render the activity of the technology itself accountable and transparent. This touches upon ethical-by-design and technical auditing (involving creation of quantitative metrics for tracing and tracking decisions, making the technologies accessible for verification and accountability). The main dimensions of technical auditing are [19]:

i. *Robustness* (Statistical Accuracy or Effectiveness): systems should be safe and secure, not vulnerable to tampering or compromising - including the data they are trained on. Key concepts in this dimension are: resilience to attack and security, fallback plan and general safety, accuracy/performance, and reliability and reproducibility.

ii. *Bias and Discrimination* (Fairness): systems should use training data and models that are free of bias, to avoid unfair treatment of certain groups. By bias we mean, for example, yielding more false positives to a group in relation to another (young people vs older people, etc.). Key sources of bias include tainted or skewed examples, limited features, sample size disparity and proxies to protected attributes.

iii. *Explainability and Interpretability*: systems should provide decisions or suggestions that can be understood by their users and developers. Key techniques in this space are individual/local explanations, population/global explanations, model-agnostic and model-specific interpretations.

iv. *Privacy*: systems should be trained following data minimization principles as well as adopt privacy-enhancing techniques to mitigate personal or critical data leakage. Key concepts in this area are data protection, quality, accuracy, integrity and access to data and decisions.

We believe that it is necessary to differentiate between auditing processes and outcomes, that is, evaluation of how the system was built (data, design, and monitoring) versus how decisions are made by the systems.

More concretely, within the context of technical auditing – there is a growing engineering expertise which can be transformed in to an extensible AI-based toolkit (i.e. test datasets, circuit-breaker software, testing Machine Learning programs) and associated framework (which includes the non-technical governance and impact assessment) for companies and researchers to ‘police’ their algorithms.

Recognising the interdisciplinary nature of such loaded concepts and challenges, systems will need to be developed as a translational tool to guide the design and engineering processes of automated

decision-making systems complete with strategies and products to ensure compliance and regulation of such systems in real-life situations. For example, we believe that such activity feeds into the ICO 'Guidelines on the AI Auditability Toolkit' [13], which can be used by developers as well as Data Scientists, Engineers and Compliance officers.

5. Monitoring Interfaces

A risk-based approach to assurance is also discussed in the literature. An important example of this is observed in the European Commission's White Paper on Artificial Intelligence [9]. Here the risk based approach is drawn from the German Data Ethics Commission [21], and two distinct notions of risk are introduced ([9] p. 17):

i. The first notion of risk is with respect to **sectors**, where high-risk is identified with respect to things such as healthcare, transport, energy, and, parts of the public sector (ex. asylum, migration, social security and employment services).

We note that all these sectors have the commonality of human impact i.e. whether a service, instruction, decision, etc. impacts on a human user and citizen. We believe that this is a broad, abstracted and blanketed approach, that is highly likely to result in two things, i. **risk aversion**, and ii. **AI will be a high cost venture**. For example, a simple healthcare booking chat bot can become economically unfeasible to develop because it falls under health. Similarly, in the context of high-risk high-reward a risk-based approach based upon sector will **discourage potentially high-positive impact AI systems** (ex. medical AI has significant risk and lifesaving potential). As such we believe this will stifle innovation (which is what the EU white paper actually calls for).

ii. The second notion of risk introduced is that 'where use means that significant risk is likely to arise (risk of injury, death or significant material or immaterial damage)'.

A concern with this categorisation of risk is that it is unclear how ***unintended consequences*** can be assessed.

Indeed, we argue that risk can be thought of in terms of known and unknown risk, and technical and non-technical risk (presented in 'Risk Matrix' Table 1 below [22]). Note that a strategy to mitigate unknown risks is discussed below in the discussion of 'Red Teaming'.

Risk Matrix		
Audit and Impact Assessment	*Known*	*Unknown*
Technical	Bias/fairness; safety; explainable; accessibility; data protection; trails; verification; comprehensibility.	Breakdown/robustness; nature of hack (theft; DOS)
Non-Technical	Governance; oversight; whistle blowing; lack of education (education/training); authorisation.	Trust; reputational; psychological and social impact; loss of skills;

Table 1: Risk Matrix outlining concerns and mitigation between technical/non-technical dimensions and known/Unknown risks [22].

Stop-Light Approach

Given the problems referred to above and the vagueness of 'risk' in these calls, drawing from industry precedence, intuitive performance dashboard stop-light interfaces have been proposed – these will facilitate monitoring of performance over time [23, 24]]. With Green, Amber and Red

representing high-performance, satisfactory-performance and poor-performance respectively. Furthermore, from a regulatory and standards standpoint, the UK's Information Commissioner's Office has a colour coded 'Assurance Rating' [25] for Data (this is reproduced (right inset). We envision that this can be used as a guidance to produce an equivalence for AI assurance.

Colour code	Internal audit opinion	Definitions
	High assurance	There is a high level of assurance that processes and procedures are in place and are delivering data protection compliance. The audit has identified only limited scope for improvement in existing arrangements and as such it is not anticipated that significant further action is required to reduce the risk of non-compliance with data protection legislation.
	Reasonable assurance	There is a reasonable level of assurance that processes and procedures are in place and are delivering data protection compliance. The audit has identified some scope for improvement in existing arrangements to reduce the risk of non-compliance with data protection legislation.
	Limited assurance	There is a limited level of assurance that processes and procedures are in place and are delivering data protection compliance. The audit has identified considerable scope for improvement in existing arrangements to reduce the risk of non-compliance with data protection legislation.
	Very limited assurance	There is a very limited level of assurance that processes and procedures are in place and are delivering data protection compliance. The audit has identified a substantial risk that the objective of data protection compliance will not be achieved. Immediate action is required to improve the control environment.

A stop light system can be used in a number of ways, including in the development phase (green, amber and red can be read in terms of ready, progressing and undeveloped respectively) and the deployment phase (with real-time monitoring being enabled, and where colours can be read in terms of how a system is performing in accordance with the purpose of its deployment).

Within the context of assurance and audit the respective colours can be read in terms of high-performing/compliant (green), low-performing/compliant (amber), non-compliant (red).

6. Unknown Risk and Red Teaming

Foundational to safety is that steps should be taken and procedures in place that *prevent harm* [26]. This preventative approach requires that risks are anticipated in order to ensure that the chances of them occurring are minimised and if they do occur then the impact is minimal. In order to do this, risk assessments are performed - in the context of the above, we can think of two kinds of risk assessment:

> **Technical audits:** these can be conducted in the development phase and for live monitoring.
>
> **Impact assessments:** these are conducted before deployment and used to put mitigation into place.

Note that in the Risk Matrix (table 1) the known technical and non-technical risks are covered by audit and impact assessment; this leaves the unknown technical and unknown non-technical risks. Within the literature one approach to addressing such unknown risks is through 'Red Teaming' [24].

Here a systematic attempt to probe, expose flaws and weaknesses in a system, process, organisation etc., both technical and non-technical, is undertaken. The 'red team exercise' assumes the persona of a hostile agent i.e. an attacker, with the hope that in exposing thereunto unanticipated weaknesses i.e. unknown risks, the risk mitigation can be improved.

Although there will still be unknown risks, through such activities it is hoped that best practice can be established; notwithstanding proprietary issues, this can be facilitated through

knowledge-transfer (via publication of methods to probe 'attack' and mitigation that has proved effective) [24].

7. Certification

Certification is the part of the assurance process by confirming that a system, process, organisation, etc. satisfies a particular standard. It is typically intertwined with regulatory requirements. However, certification can also be granted by industry bodies or other recognised authorities. We read certification as a final 'stamp' or confirmation, which can be achieved via providing evidence and proving that a system, process, organisation has satisfied the set standards.

Certification may come in a number of forms, including:

> **Certification of a system:** here, likely to align with national regulatory and standard bodies, the use of AI i.e. the systems and governance, may be certified as trustworthy or responsible. This may be akin to the granting of an organisational licence.
>
> **Sector Specific Certification:** here it is possible that sector standard bodies and regulators issue their own sector specific certification.
>
> **Certification of a Responsible Agent:** good practice and industry standards within the context of data protection has led to the position/role of a 'Data Protection Officer', and by analogy something akin to a 'Responsible AI Officer' may emerge. These officers may be certified.
>
> **Certification of AI Engineers:** Here the AI engineers may be certified, as though being granted a licence or admission into an an accreditation organization (c.f. British Medical Association).

Another possibility is that certification may be issued for specific aspects of a system; here certifications for **robustness**, **explainability**, **privacy** and **fairness** may be issued.

8. Insurance

Closely related to assurance is the insurance of AI systems. It is possible that this will become a significant risk mitigation requirement for companies engaged in automation, and as such a significant market for insurers. We envision that this will align closely with explainability and AI auditing in accordance with regulation and compliance that satisfies trustworthy AI standards.

9. Conclusion

In this article we have outlined a number of themes that we anticipate may emerge as a result of the maturation of the assurance debate. Indeed, some of these themes have already reached a point where there is considerable literature - this is particularly true of the technical assessment theme (referred to often in the literature and regulatory debates as 'AI auditing'). Conversely, some of these themes have as yet generated less attention - insurance of AI systems is an example of this.

We believe that the moment of crystallisation, with respect to assurance of AI systems, will come when the national and international regulatory and standards bodies decide upon the their respective approaches; the key issue will be how 'hard' (highly perspective, novel regulation) or 'soft' (principles approach, application of - or updating/amending of - already existing regulation) the intervention will be.

10. References

[1] Schulte, J. AI-assisted recruitment is biased. Here's how to make it more fair. World Economic Forum (May 2019). https://www.weforum.org/agenda/2019/05/ai-assisted-recruitment-is-biased-heres-how-to-beat-it/ (Accessed: 30 August 2020).

[2] Hao, K. AI is sending people to jail—and getting it wrong. MIT Technology Review (19 January 2019) https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/ (Accessed: 30 August 2020).

[3] Burkell, J., and Priscilla M. Regan, P. M. Voter preferences, voter manipulation, voter analytics: policy options for less surveillance and more autonomy.Internet Policy Review 8, no. 4 (2019), 1-24.

[4] Challen, R., et al. Artificial intelligence, bias and clinical safety. BMJ Quality & Safety, 28(3). (2019), 231-237.

[5] Hleg, A. I. Ethics guidelines for trustworthy AI. B-1049 Brussels (2019).

[6] Piano, S. L., 2020. Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. Humanities and Social Sciences Communications, 7(1), pp.1-7.

[7] Leslie, D. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. (2019).

[8] Parens, E. Technology as Value-Free and as Value-Laden. Oxford University Press, In book: Shaping Our Selves. (September 2014), 71-92.

[9] 'White Paper on Artificial Intelligence: A European approach to excellence and trust', European Commission (February 2020).

[10] Kazim, Emre and Koshiyama, Adriano, Lack of Vision: A Comment on the EU's White Paper on Artificial Intelligence (March 20, 2020). Available at SSRN: https://ssrn.com/abstract=3558279 or http://dx.doi.org/10.2139/ssrn.3558279

[11] Artificial Intelligence and Public Standards: report. Committee on Standards in Public Life, Chair, Lord Evans of Weardale KCB DL (10 February 2020).

[12] Kazim, E., and Koshiyama, A. No AI Regulator: An Analysis of Artificial Intelligence and Public Standards Report (UK Government) (February 26, 2020). Available at SSRN: https://ssrn.com/abstract=3544871 or http://dx.doi.org/10.2139/ssrn.3544871.

[13] Guidance on the AI auditing framework Draft guidance for consultation. Information Commissioner's Office (February 2020). https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-consultation-on-the-draft-ai-auditing-framework-guidance-for-organisations/ (Accessed: 30 August 2020).

[14] Kazim, E., and Koshiyama, A. A Review of the ICO's Draft Guidance on the AI Auditing Framework (May 12 2020). Available at SSRN: https://ssrn.com/abstract=3599226 or http://dx.doi.org/10.2139/ssrn.3599226.

[15] Explaining decisions made with AI. Information Commissioner's Office & The Alan Turing Institute (2020)

https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/ (Accessed: 30 August 2020).

[16] Kazim, E., and Koshiyama, A. Explaining Decisions Made With AI: A Review of the Co-Badged Guidance by the ICO and the Turing Institute (July 20, 2020). Available at SSRN: https://ssrn.com/abstract=3656269 or http://dx.doi.org/10.2139/ssrn.3656269.

[17] Mueller, H. and Ostmann, F. 'AI transparency in financial services – why, what, who and when?', Financial Conduct Authority (19 February 2020). https://www.fca.org.uk/insight/ai-transparency-financial-services-why-what-who-and-when (Accessed: 30 August 2020).

[18] Interim report: Review into bias in algorithmic decision-making. Centre for Data Ethics and Innovation (July 2020). https://www.gov.uk/government/publications/interim-reports-from-the-centre-for-data-ethics-and-innovation/interim-report-review-into-bias-in-algorithmic-decision-making (Accessed: 30 August 2020).

[19] Koshiyama, A. and Engin, Z., 2019. Algorithmic Impact Assessment: Fairness, Robustness and Explainability in Automated Decision-Making (June 2019). https://zenodo.org/record/3361708#.X0u4sOhKhPY (Accessed: 30 August 2020).

[20] McGregor, L., Murray, D., and Ng, V. (2019). International human rights law as a framework for algorithmic accountability. International & Comparative Law Quarterly, 68(2), 309-343.

[21] German Data Ethics Commission. (2019). Opinion of the Data Ethics Commission. Bmjv.de. Available at: www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.pdf (Accessed: 30 August 2020).

[22] Kazim, E., and Koshiyama, A. Human Centric AI: A Comment on the IEEE's Ethically Aligned Design (April 13, 2020). Available at SSRN: https://ssrn.com/abstract=3575140 or http://dx.doi.org/10.2139/ssrn.3575140.

[23] A. Bellotti, D. J. Hand, S. Khan, Predicting Through a Crisis. Second White Paper. Validate AI (July, 2020).

[24] Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., ... & Maharaj, T. (2020). Toward trustworthy AI development: mechanisms for supporting verifiable claims. arXiv preprint arXiv:2004.07213.

[25] Assurance Ratings. Information Commissioner's Office. https://ico.org.uk/for-organisations/audits/ (Accessed 30 August 2020)

[26] Kazim, E., and Koshiyama, A. A High-Level Overview of AI Ethics (May 24, 2020). Available at SSRN: https://ssrn.com/abstract=3609292 or http://dx.doi.org/10.2139/ssrn.3609292.

[27] Kazim, E., and Koshiyama, A. The Interrelation Between Data and AI Ethics in the Context of Impact Assessments (June 15, 2020). Available at SSRN: https://ssrn.com/abstract=3627561 or http://dx.doi.org/10.2139/ssrn.3627561.