ASTL Algorithm Standards & Technology Lab.

Towards AI Standards

Thought-leadership in AI legal, ethical and safety specifications through experimentation

UCL Standards Lab Joint Publication ●

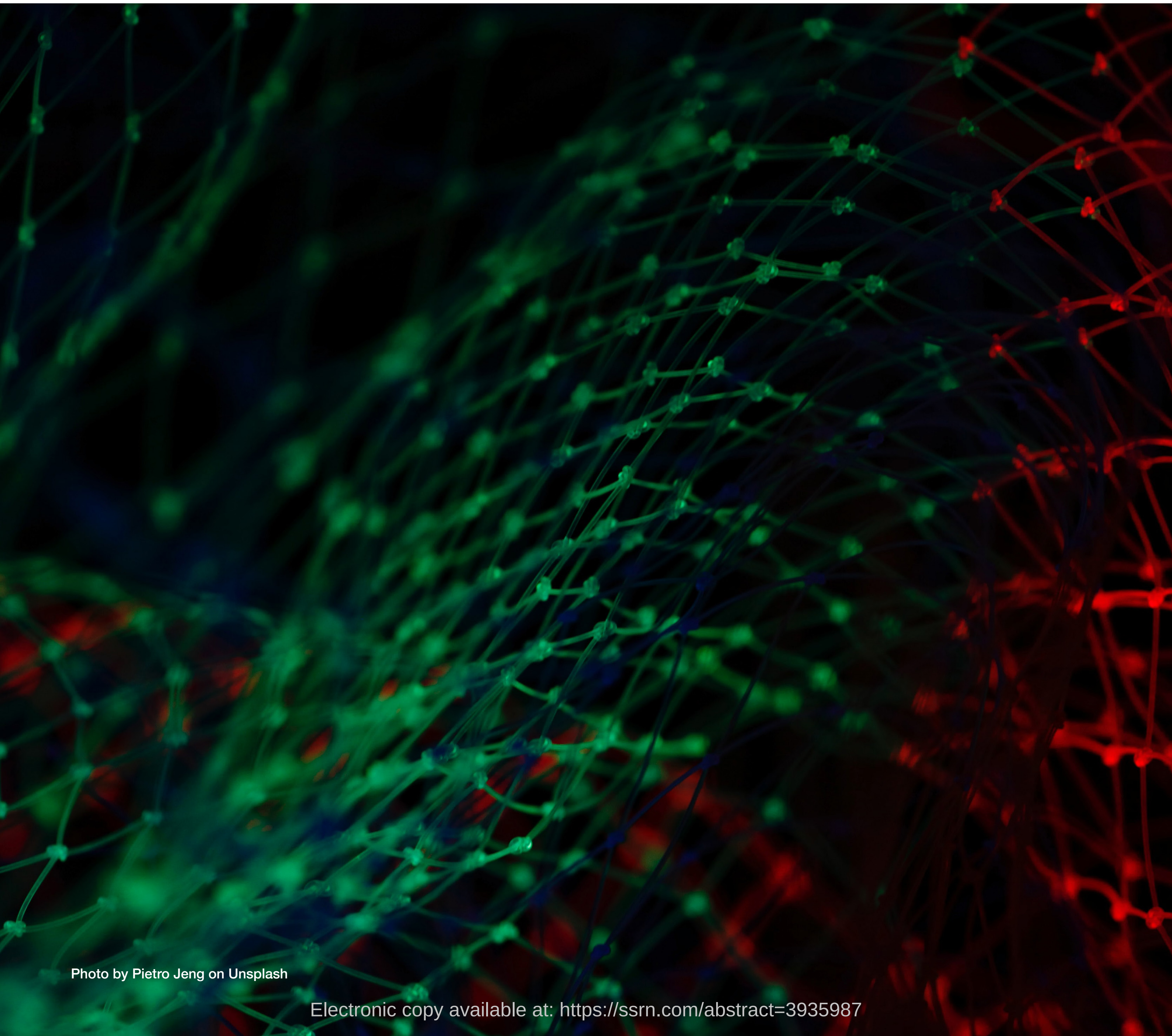


Photo by Pietro Jeng on Unsplash

Abstract

With the rapid adoption of algorithms in business and society there is a growing concern to safeguard the public interest. Researchers, policy-makers and industry sharing this view convened to collectively identify future areas of focus in order to advance AI standards - in particular the acute need to ensure standard suggestions are practical and empirically informed. This discussion occurred in the context of the creation of a lab at UCL with these concerns in mind (currently dubbed as *UCL The Algorithms Standards and Technology Lab*). Via a series of panels, with the main stakeholders, three themes emerged, namely (i) Building public trust, (ii) Accountability and Operationalisation, and (iii) Experimentation. In order to forward the themes, lab activities will fall under three streams - experimentation, community building and communication. The Lab's mission is to provide thought-leadership in AI standards through experimentation.

Keywords

Artificial Intelligence

Ethics

Risk-Management

Robustness

Privacy

Fairness

Explainability

Auditability

Risk

Assessment

Authors:[*$]

Roseline Polle[$,1], Emre Kazim[$,1,2], Graca Carvalho[$,1] Adriano Koshiyama[$1,2], Catherine Inness[1], Andrew Knight[3], Caroline Gorski[4,5], David Barber[1], Elizabeth Lomas[1], Emine Yilmaz[1], Gary Thompson[6], Ghazi Ahamat[7], Giles Pavey[8], Katie Platts[8], Lukasz Szpruch[9,10], Miro Gregorovic[11], Miguel Rodrigues[1], Pamela Ugwudike[12], Parashkev Nachev[1], Pete Rai[13], Philip Treleaven[1], Randy Goebel[14]

[1]University College of London, [2]Holistic AI, [3]RICS, [4]Rolls-Royce, [5]Emergent Alliance ,[6]Big Innovation Centre, [7]Centre For Data Ethics and Innovation, [8]Unilever, [9]Unversity of Edinburgh, [10]The Alan Turing Institute, [11]London Stock Exchange, [12]University of Southampton, [13]Cisco Systems, [14]University of Alberta

[$] *Corresponding authors: [roseline.polle.19, e.kazim, adriano.koshiyama.15]@ucl.ac.uk*

[*] *Listed authors are those who contributed substantive ideas and/or work to this report. Contributions include writing, research, and/or review for one or more sections, and/or via ongoing discussions. As such, except for the corresponding authors, inclusion as author does not imply endorsement of all aspects of the report; aside from the first five, authors are listed in alphabetical order. The list of authors may change, subject to approval from their respective organizations.*

Table of Contents

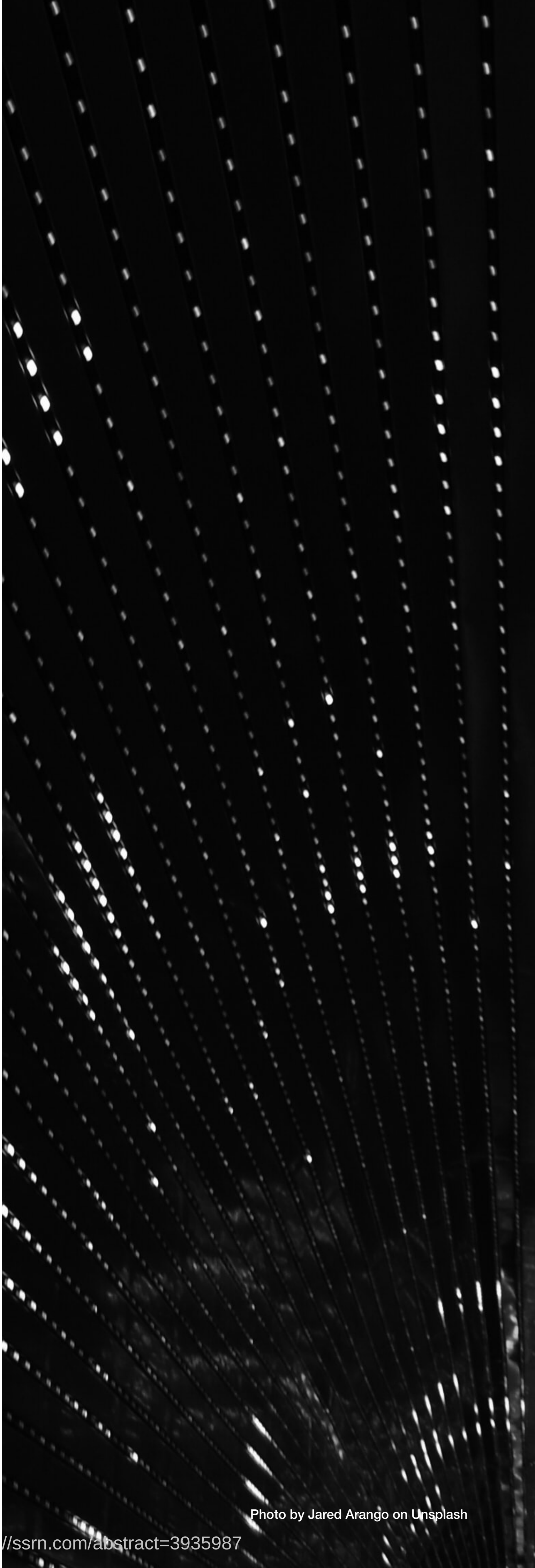


Photo by Jared Arango on Unsplash

EXECUTIVE SUMMARY

Motivation
Algorithms are rapidly becoming pivotal in business and society. There is a growing concern to safeguard the public interest. As such, deploying algorithms requires good governance. To this, standards and regulations beyond GDPR are being proposed. Researchers, policy-makers and industry sharing this view convened to address these issues - in particular the acute need to ensure standard suggestions are practical and empirically informed. This discussion occurred in the context of creating a lab at UCL with these concerns in mind *(The Algorithms Standards and Technology Lab).*

Workshop
This workshop was designed as a series of short panels with the main stakeholders interested in playing a role in the Lab. Four sector focused panels were held (names listed are those comprising the panel):

- UCL Internal Computer Science: Prof Philip Treleaven (Financial Computing), Prof Emine Yilmaz (Computer Science), and Prof David Barber (AI Centre).

- UCL Cross-Departmental: Prof Parashkev Nachev (Institute of Neurology), Prof Miguel Rodrigues (Electric Engineering), Dr Elizabeth Lomas (Information studies).

- External engagement: Dr Lukasz Szpruch (The Alan Turing Institute), Caroline Gorski (Rolls-Royce), Ghazi Ahamat (Centre for Data Ethics and Innovation).

- Industry partners: Pete Rai (Cisco), Katherine Platts (Unilever), and Mick Gregorovich (LSEG)

Emergent Themes
After collating points raised in the panels, discussion and feedback, the following agenda points were found:

- Building public trust: addressing lack of public trust, lack of common language, and facilitation of communications.

- Accountability and Operationalisation: need for clear and public standards, transparency, explainability and accountability mechanisms across all stakeholders, respect for sector specificity, develop industry best practice, and need to consider global scope of AI technologies.

- Experimentation: empirically informing standards and public and policy debate, safe spaces to facilitate knowledge exchange, avoiding silos and initiating joint academia-industry pilots.

Activities
In order to address the themes, lab activities will fall under three streams experimentation, community building and communication.

Lab Mission

“Thought-leadership in AI standards through experimentation”

INTRODUCTION

Algorithms are rapidly becoming pivotal in business and society. There is a growing concern to safeguard the public interest. As such, deploying algorithms requires good governance. To this, standards and regulations beyond GDPR are being proposed (e.g. European Commission, 2021; UK Committee on Standards in Public Life, 2020). Researchers, policy-makers and industry sharing this view convened to address these issues - in particular the acute need to ensure standard suggestions are practical and empirically informed. This discussion occurred in the context of creating a lab at UCL with these concerns in mind *(The Algorithms Standards and Technology Lab).*

THEMATIC STAKEHOLDER PANELS

The half-day event was structured in the form of a workshop. This workshop was designed as a series of short panels with the main stakeholders interested in playing a role in the Lab. Following the opening, four sector focused panels were held, chaired by Graca Carvalho and Dr Emre Kazim (UCL Computer Science). Names listed are those comprising the contributions panel, with their key points drawn out:

Introduction:

● Dr Adriano Koshiyama (UCL Computer Science): ***Evolving landscape*** - Increasing amount of AI being developed; Changing regulatory and operational environment; Translating AI ethics principles into practice; Concern with financial and reputational damage; Void of technology standards for AI; (See Koshiyama et al., 2021; Kazim and Koshiyama, 2020a, 2020b).
Ambition - Be the space for informed and empirically evaluated debates and technical solutions on standards; Leverage our global network of experts to inspire and design sound policies and practices; Build a relationship model that can lead to experimentation, translational research, and best practices.

● Prof Randy Goebel (University of Alberta, Vice Provost): Need to build a trajectory and smooth transition from research to deployment in industry; Demand for trustable systems for AI has never been higher; there is a key role for key and trusted institutions to smooth the path. (See Goebel et al. 2018; Longo, Goebel et al., 2020).

UCL Internal Computer Science:

● Prof Philip Treleaven (UCL Financial Computing): Global interest; run projects with and invite contributions from student teams worldwide; should help generate these projects; students as an interface to companies; experimentation. (See Treleaven et al., 2017, 2019).

● Prof Emine Yilmaz (UCL Computer Science): Need to set up standards (AI for decision making); metrics for fairness; independent and reliable datasets (test sets); Industry awareness; academia makes is uniquely placed to release large training datasets; build AI that can work together with humans; Explainability. (See Yilmaz et al., 2020).

● Prof David Barber (UCL AI Centre): Access to data; not losing sight of the enormous transformation potential; we need standards for AI that companies will soon be importing (from other jurisdictions); deliver meaningful impact as quickly and safely as we can. See (Barber, 2012, 2020).

UCL Cross-Departmental:

● Prof Parashkev Nachev (Institute of Neurology): Standards require must be some kind of measurement/ quantification; fidelity of the model; show how the current approach to modelling in healthcare is deficient or unfair; show how AI and more complex SoTA models are better; create metrics of fidelity that are more appropriate to the individual case; measure causality, generalisability, robustness of distribution of training data; value-based considerations - 'quantitative ethics'; Benchmarking. (See Nachev et al., 2008.)

● Prof Miguel Rodrigues (Electric Engineering): Need to look at it holistically; We do not fully understand SoTA algorithms. (See Rodrigues et al., 2003)

● Dr Elizabeth Lomas (Information studies): Pipeline documentation; experiment with human-in-the-loop; explainability to the general public; explainability to government; more academic access to data for societal purposes; accountability; language is important; narratives around the metrics; Need to work with public engagement and the media. (See Lomas, 2010, 2020).

External engagement:

● Dr Lukasz Szpruch (The Alan Turing Institute): Translate principles into practical tools; bottom-up approach; standardised data sets and benchmarks; backtesting; trade-offs. Will be clearer from bottom-up approaches; Robustness is the key; Robustness of fairness, explanations. certification guarantees, including stress testing; Guarantee that the model performs well under a specific environment. (See Cohen, 2021).

● Caroline Gorski (Rolls-Royce): Lack of fully explainable AI cannot hold industry back; Regulation is crucial to enabling safe economic growth of the industry; access to data, negative narrative, quality standards and metrics; more responsible data stewards.

● Ghazi Ahamat (Centre for Data Ethics and Innovation): We have moved past the concept of regulation vs innovation; Uncertain trust in AI will hold back adoption; interdisciplinary information sharing; Data sharing; We do not need a single concept of fairness or risk to make progress; need more general language and a common understanding between developers, accountable executive, regulators and the affected individual. (See UK-CDEI, 2020).

Industry partners:

● Pete Rai (Cisco): Practical steps that engineers can take; robust terminology; operationalise; as robust and compulsory as their security process; safe forum; globally not regionally. Produce something that has been designed for regulators to rubber stamp; quickly and iteratively; separate exercise to policing what bad actors could do; companies coming together who want to do it properly.

● Katherine Platts (Unilever): Things are going to go faster; globally set principles then tailored to local regulations; restore the trust in digital society so we can innovate and explore the advantages; adaptable; Building skills; ethics by design.

● Mick Gregorovich (LSEG): Cultural alignment; proactive; language; buy-in and general understanding; Education; internal and customers.

Feedback was also submitted from participants, which we similarly draw out keywords from below:

● Dr Pamela Ugwudike (University of Southampton): AI transparency and accountability; Limited access to relevant data; climate of AI demonisation; upsides of AI should be emphasised as well; develop gold standard metrics for fairness; offer student projects and engage with students from other countries via virtual internships; How do we assess the utility of AI systems; current ethical standards focus narrowly on data-related problems; multidimensional standards should be developed; in terms of AI infrastructure, standards are necessary e.g., data collection processes etc.; validating inputs and outputs; how can the fairness of a model be measured if subgroups are not included in the model? (See Ugwudike, 2020).

● Gary Thomson (Big Innovation Centre): Issue of ethical and safety vs scalability; audit automated to give speed to the adoption across enterprises.

● Andrew Knight (RICS): Very keen to see the market understand that we need to go beyond GDPR, and other jurisdictional legislation; the development of a common language for the whole AI space; the need to cater for horizontal and vertical applications both of which will be developed in many cases for international use. One area that the report could highlight is the interest and role that standard setters and regulators like RICS need to perform to align with the proposed approach to both support the use of algorithms in their internal functions and how they regulate members that use AI etc. Andrew is actively leading an internal project to look at these issues from the perspective of automated valuation models (AVMs).

● **Additional points raised:** suggested points raised in Bank of England AI Public Private Forum: Incremental standards may be needed for most elements of traditional data Governance (data documentation, quality, retention, privacy, sovereignty, and security); challenges around the adoption of standards (differential privacy, federated learning, homomorphic encryption, personal data wallets and federated consent management); Equal access and fair pricing of third-party data; pricing of third-party data based on the value such data generates. Ensuring data regulators and financial service regulators stay aligned; Data auditing, certification and attestation; Lack of certification for data sets; model auditing holistically; How to fit any new AI auditing framework into an organisation's existing; enterprise-wide risk management frameworks; international level; Industry-agnostic coordination; any guidance should not be overly prescriptive and should provide case studies; Creation and use of cross-organisational datasets that can support the public good and increase financial inclusion; Cooperating on 'data for good' is extremely valuable.

THEMES

After collating points raised in the panels, discussion and feedback, the following agenda points were found -

- Building public trust: addressing lack of public trust, lack of common language, and facilitation of communications.

- Accountability and Operationalisation: need for clear and public standards, accountability mechanisms, respect for sector specificity, develop industry best practice, and need to consider global scope of AI technologies.

- Experimentation: empirically informing standards debate, safe spaces to facilitate knowledge exchange, avoiding silos and initiating joint academia-industry pilots.

Below we expand on each theme by fleshing out the core points of discussion.

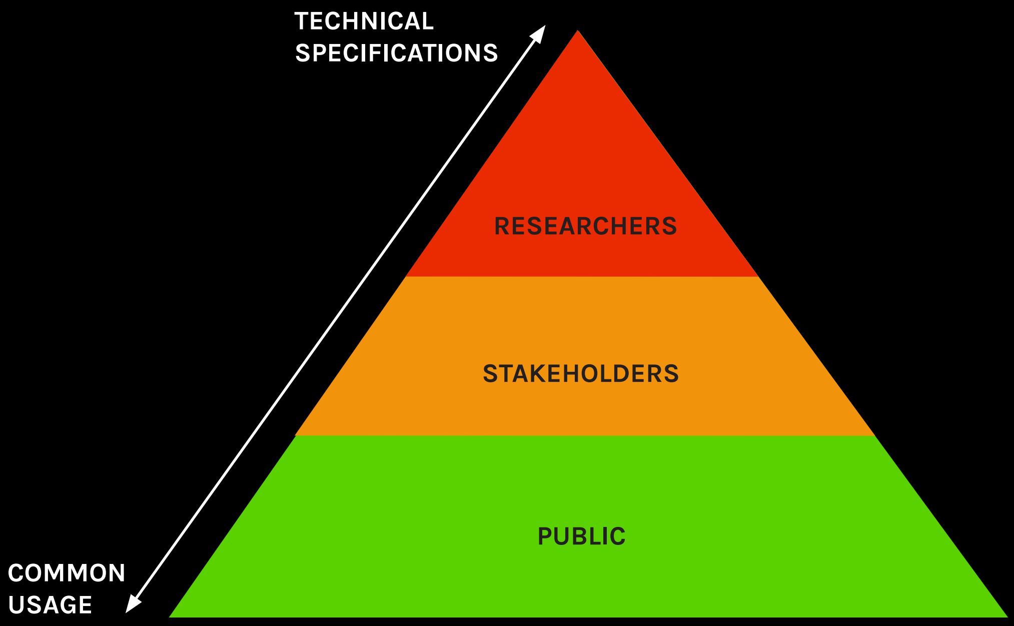
Photo by Esaias Tan on Unsplash

Theme 1. Building public trust

Public concern:
A common observation amongst participants is that there is a lack of trust by the public in AI and data technologies, and that such concern hampers progress i.e. development and adoption of such systems. While concern for accountability is shared by all participants i.e. systems must be governed well and responsibly, there is the danger of elevated public concern by inadvertently foregoing potential benefits to all. As such, there is a need to restore trust and highlight the capability of AI to improve processes in particular in relative terms such as via a comparison with respect to current processes (services, products etc. i.e. in terms of performance but also fairness, explainability and robustness).

Clarity in Language & Communication:
The need for clarity in language (for example in terms such as 'AI', 'explainability', 'fairness' etc...) was expressed with respect to three layers of communication: the public, organisational stakeholders (developers, accountable executives, regulators and the affected individuals), and finally researchers. Terms used, although they will have more technical granularity to them, can nonetheless signify the common concepts. The importance of developing coherent narratives should also be encouraged in order for the public and non-experts to understand and debate AI without technical expertise. For example, narratives could be developed around things such as commonly used metrics (and why these are used) and the science/ engineering of these technologies. Additionally, thought should be given to communications strategies, including mechanisms of knowledge exchange and forums for discussion. These ideas are summarised with the pyramid in Figure 1, with a common language needed horizontally for effective communication within each level, but also vertically so concepts are understood by everyone.

FIGURE 1:
Building a common language between (i) Researchers (academic working on these subjects), (ii) Organisational stakeholders (developer, regulator, NGO, accountable executive, etc,..), (iii) Public (individuals and media).



Theme 2. Accountability and Operationalisation

Need for standards:
The strong view from participants is that there is an industrial need for clear, public standards and regulation – this is a view shared by the academy i.e. that more accountability is needed for industry (developers and deployers of systems) and those impacted by algorithms more widely. Such accountability was spoken of in terms of standards that may be proposed in a number of areas, including a template for pipeline documentation, a standard set of metrics when appropriate (e.g. fairness), the use of independent and reliable datasets (for both training and test sets to evaluate model against), and an independent certification guarantee/service that the model performs well under a specific environment/ conditions and tested documentation samples explaining algorithmic impacts to the wider public. Crucially, the view was that these standards could be developed through projects that serve as examples, ensuring the standards are empirically informed and practica**l.**

There was a call to recognise sector and domain differences:
In doing so, such differences should be reflected in the nature of what the standards are for those sectors and domains. For example, fairness metrics in the context of recruitment and labour more generally, measuring the fidelity of models for healthcare, safety provision for car and aeronautic industry, etc… There was an open question regarding how prescriptive sector and domain standards should be. For instance, there was a view that general standards should be proposed that can be adaptable with respect to the standards and regulation for these context dependent cases.

Industry cannot wait:
The development and deployment of algorithms is already well underway – indeed this is rapidly growing – as such those who develop and deploy cannot wait for regulation and standards to be determined but must act now. Industry has started to already act, and is doing this in the context of developing their own practices and implementing good governance. The view is that this work will continue in parallel with regulatory and standards interventions, which it is hoped will cohere with best practice (of those with well developed governance). Indeed, the view is expressed that work on standards and regulatory proposals needs to be accelerated.

Longer term risk:
Often risk in the context of algorithms is discussed in terms of the use case, and indeed, as noted above, the view is that risk is context specific, however concern with longer term risk was also raised. Here the example of biological computing was explored, and the view that transparency and responsibility should be in mind at research/development inception phase (c.f. vaccine development).

Global Scope:
That algorithmic systems are seldom developed and deployed solely in one jurisdiction was raised in the context of ensuring that any proposals, standards suggestions etc. should bear in mind the global reach, applicability and scope of these systems.

Theme 3. Experimentation

Evidence Approach:
Participants identified the need for an interdisciplinary experimental space, where translation of ethics principle into practice can be freely explored. In addition to enabling the identification of potential issues and solutions for operationalisation of data and AI ethics practices, this space would enable transfer of knowledge, expertise and skills between different actors. Two mechanism for achieving this were proposed:

• Sandboxing**:** This would include a safe space where industrial actors can share progress on solutions for trustworthy AI without reputational risks or IP limitations, including documentation i.e. testing the documentation with tech regulators but in addition statements on the algorithms and their impact with sample citizens, and industry offering sharing 'half-world' perspectives for on work in progress, for discussion. This serves as a transfer knowledge platform as well as an experimental place where data can be shared freely. There are some legal and reputational challenges associated with the creation of such a structure (e.x. IP/ competition/privacy) which will need to be addressed. Crucially this will entail a mechanism for industry and government partners who are granting access to data for such experimentation.

• Pilot programmes: The platform could also serve as a space for experimentation via small pilots and other projects by researchers, notably in collaboration with industry. The platform should make use of students' resources for this purpose. This bottom-up approach will allow for the identification of the problems specific industries might encounter in practice during operationalisation of ethics principles. It will help industries develop their own set of ethics practices as well as inform governments on potential regulation.

Diversity of participants:
Comparatively to the more concrete space of standards and regulations, there is the subjectivity of ethics and regional practices. As such, it is essential that as a policy diversity in terms of approaches to projects is ensured. Global collaboration is essential. The platform would be a powerful tool in this respect, as well as a privileged space for transfer knowledge between academia and industries, including multidisciplinarity from different academic units.

The table below summarises the main points from each theme, as well as the questions to explore that were derived from these.

	Problem statement	Questions
Theme 1: Building public trust	● Lack of public trust ● Lack of common language ● No standard mechanism of communication ● Narratives for non-experts	● How can we engage with the public in order to develop trust? ● How can we develop a common vocabulary? ● How can we meaningfully gauge public input? ● Can we be creative in our narrative telling?
Theme 2: Accountability & Operationalisation	● Need for clear and public standards, and accountability mechanisms (respecting sector specificity) ● Need enable industry developing good practice in absence of regulation ● Longer term risk view ● Global nature of AI means that national view is limited	● How can we accelerate standards? ● How can we share and utilize industry best practice? ● How can we communicate/facilitate standards suggestions with regulators (national/International)? ● How can we communicate and explain algorithms and their beneficial impacts to the public?
Theme 3: Experimentation	● Current standards are not empirically informed - problems of operationalising ● Need for safe space to share partial solutions ● Need for experimentation via the running of pilots - including industry and academia	● Can we ensure that standard suggestions are empirically informed/practical? ● Can we foster a safe space for experimentation? ● How can we foster knowledge, expertise and skills exchange?

LAB ROADMAP

Activities

The Lab's activities will focus on addressing the areas identified within the themes noted above. They are articulated around three main pillars: Experimentation, Communication and Community Building. **Figure 2** presents an overview of these for the initial programme of activities.

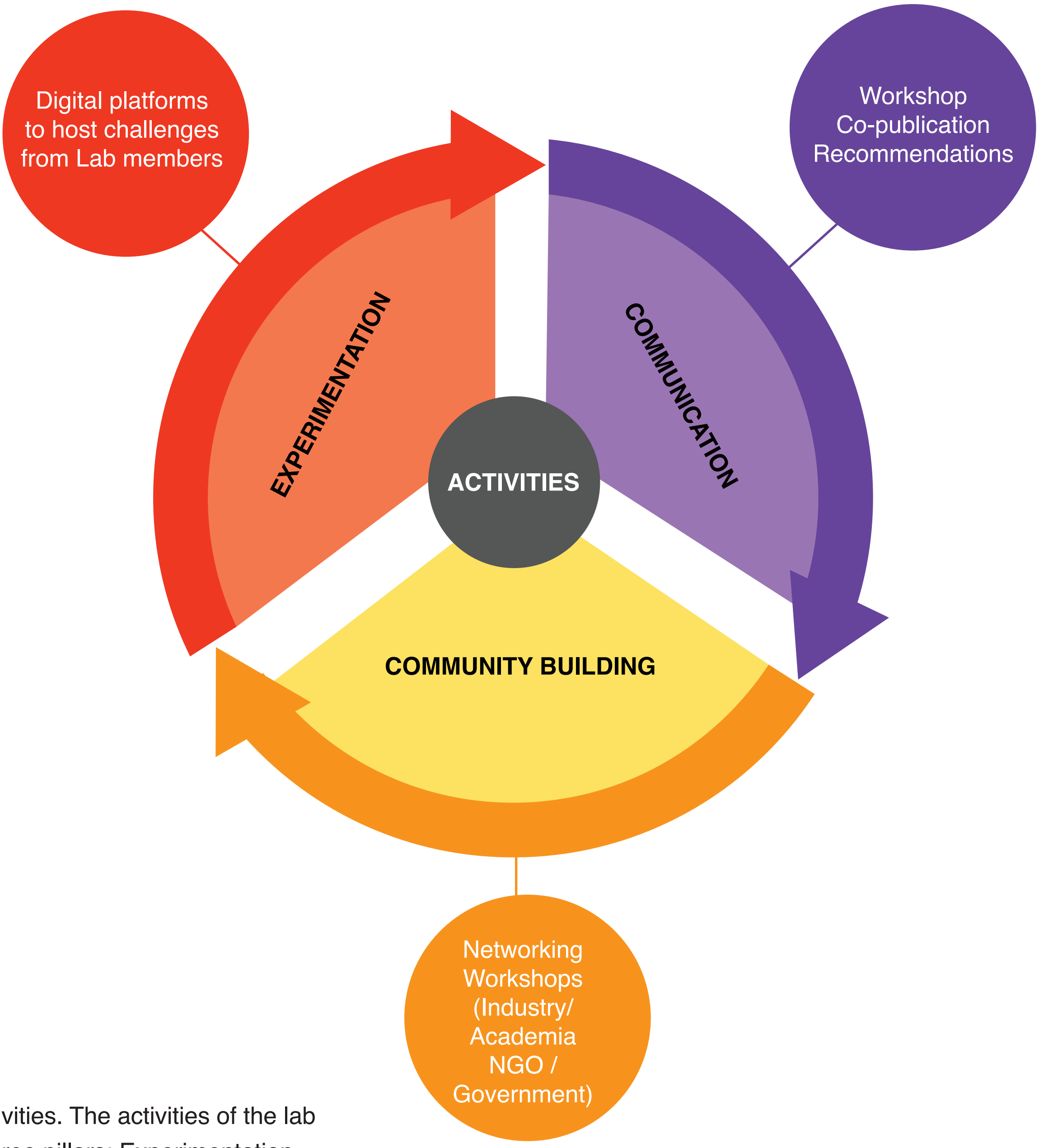

Digital platforms to host challenges from Lab members
Workshop Co-publication Recommendations
EXPERIMENTATION
COMMUNICATION
ACTIVITIES
COMMUNITY BUILDING
Networking Workshops (Industry/ Academia NGO / Government)

FIGURE 2:
Initial programme of activities. The activities of the lab are structured around three pillars: Experimentation, Communication and Community Building. Each pillar feeds and informs the others. Practical actions by the lab include setting up the required infrastructure to enable experimentation (i.e. digital platform), organising networking events to bring the industry, government, NGOs and academia together, and communicate with the different stakeholders (including the public).

Challenges

Drawing upon the literature and core themes that have emerged in the space of AI ethics, we have identified five core challenges (accountability, robustness, bias, explainability and privacy - expanded upon below). These challenges form the core research verticals encompassing the field of trustworthy AI. We envision that by framing the experimentation, recommendations and communications of the lab with these research verticals, the above identified themes of the lab will be addressed.

These challenges are divided into two broad streams, namely accountability/governance and assessment verticals (see Koshiyama et al, 2021):

● **Accountability:** 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. Where appropriate, analysis will survey governance norms and sector specific particularities in the context of ML/AI development and product deployment.

● **Assessment verticals:** 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/verticals of technical governance are:

o **Robustness (Statistical Accuracy or Effectiveness):** quality of a system to be safe, not vulnerable to tampering.

o **Bias and Discrimination (Fairness):** quality of a system to avoid unfair treatment of individuals or organizations.

o **Explainability and Interpretability:** quality of a system to provide decisions or suggestions that can be understood by their users and developers.

o **Privacy:** quality of a system to mitigate personal or critical data leakage.

Figure 3 shows how the themes, challenges and outputs connect together.

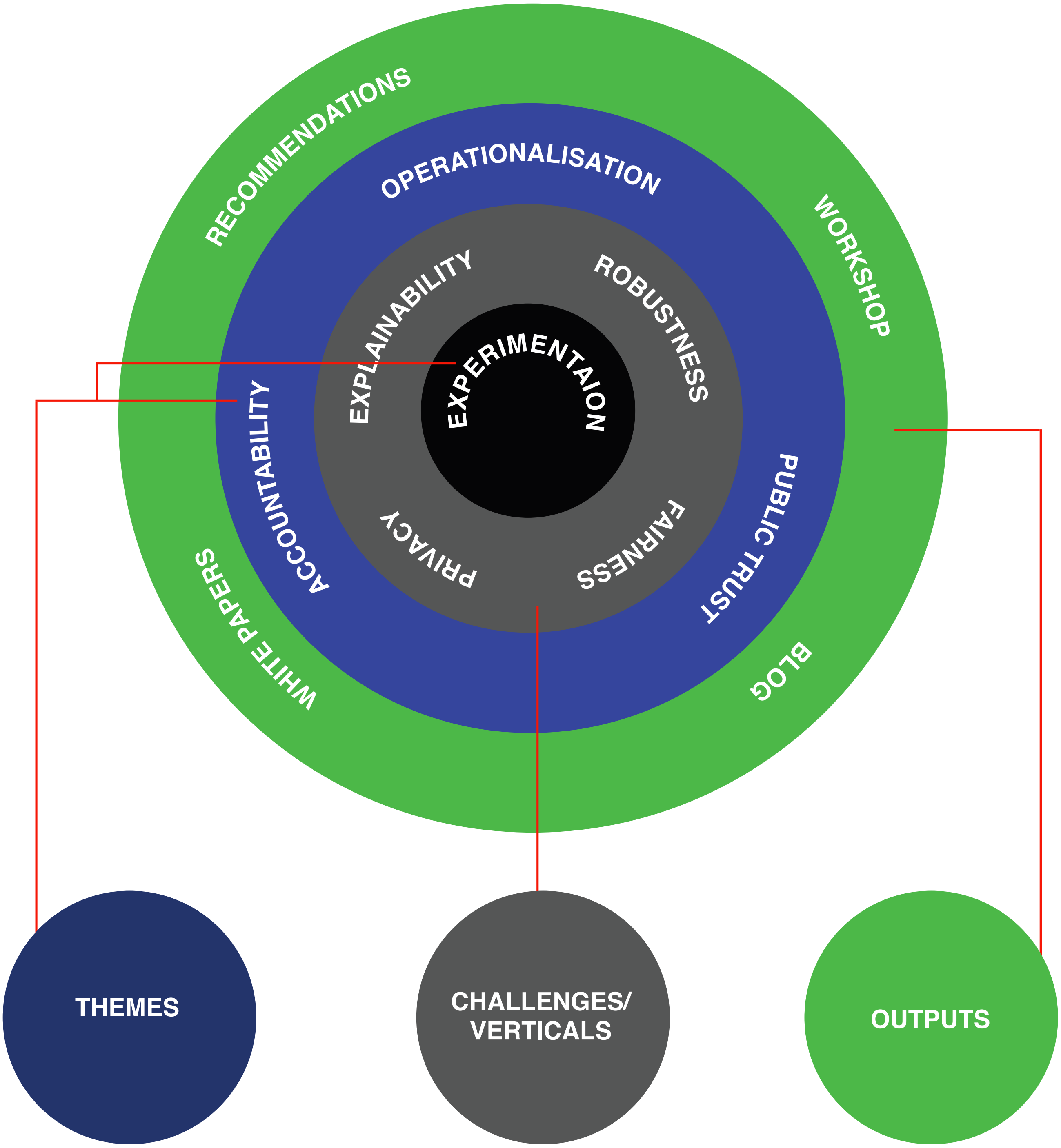


FIGURE 3:
Themes, challenges and outputs. At the core of the lab is experimentation and the need to empirically inform standards and the public and policy debate. Experimentation is framed across the four assessment verticals (Explainability, Robustness, Fairness and Privacy). Defined as a theme of the lab in itself, it feeds into the remaining themes (Accountability and Operationalisation, and Public trust). The lab will notably communicate findings to the different stakeholders (including the public) by means of workshops, blog, or white papers.

Outputs

The lab will publish a number of white papers. These will include position papers - such as this one - gathering views from stakeholders, as well as papers based on internal research. Pilots will be documented as and when projects occur. From the different streams, recommendation papers will also be published. Public facing communication channels, such as a blog or newsletter, will also be open source. Finally, the lab intends to create a digital platform to hold datasets and models from partners along with challenges to solve. Initial thoughts are to use Kaggle or existing regulatory sandboxes to host these. These challenges will be accessible to the wider public to experiment with solutions.

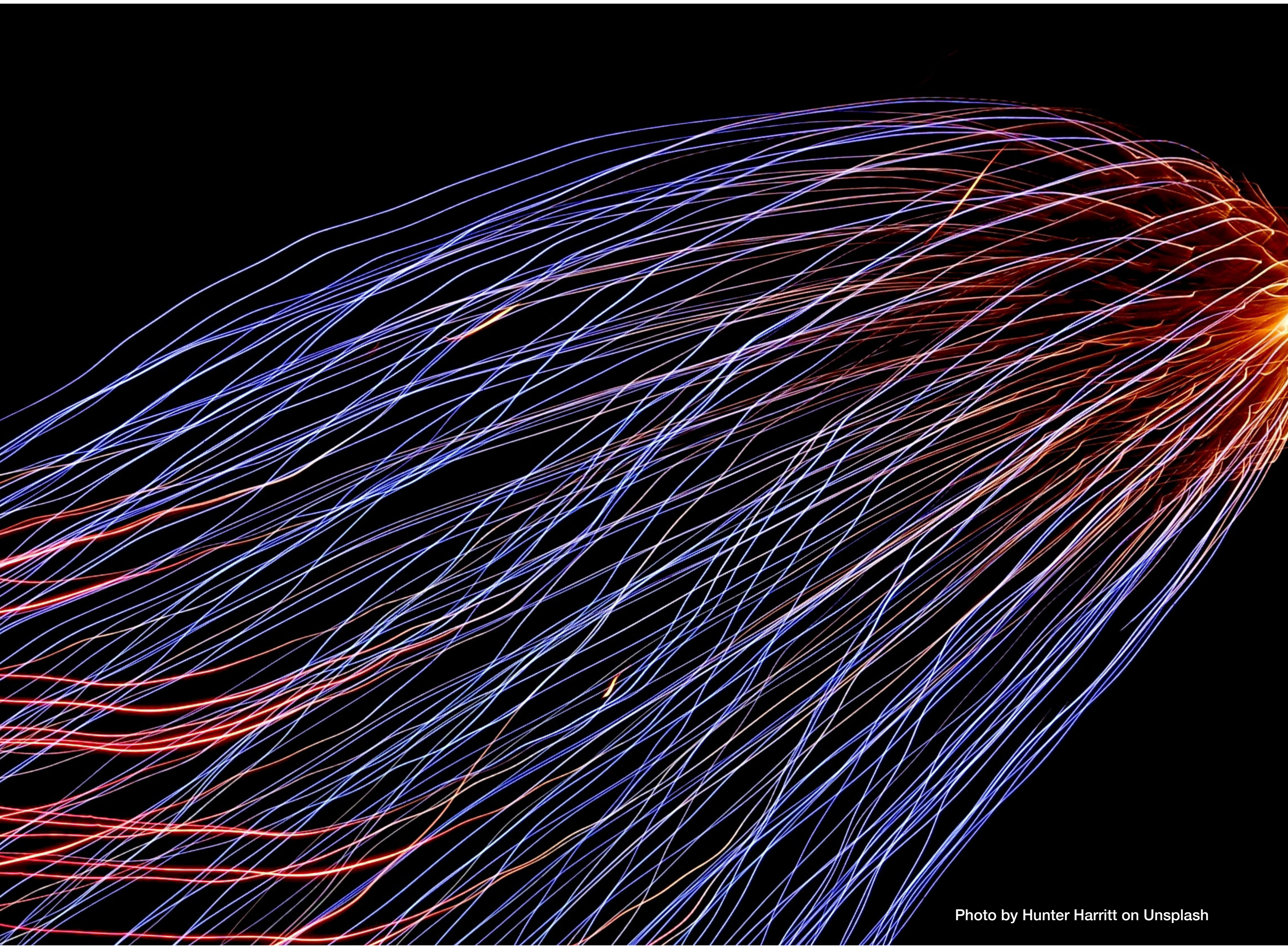


Photo by Hunter Harritt on Unsplash

REFERENCES

Ada Lovelace Institute. (2021). Themes. https://www.adalovelaceinstitute.org/our-work/themes/

AI Now Institute. (2021). "A New AI Lexicon: Responses and Challenges to the Critical AI discourse".https://medium.com/a-new-ai-lexicon/a-new-ai-lexicon-responses-and-challenges-to-the-critical-ai-discourse-f2275989fa62

Algorithmic Accountability Act of 2019, H.R.2231, 116th Cong. (2019). https://www.congress.gov/bill/116th-congress/house-bill/2231

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics surveys, 4, 40-79.

Arnold, M., Bellamy, R. K., Hind, M., Houde, S., Mehta, S., Mojsilović, A., ... & Varshney, K. R. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM Journal of Research and Development, 63(4/5), 6-1.

Arya, V., Bellamy, R. K., Chen, P. Y., Dhurandhar, A., Hind, M., Hoffman, S. C., ... & Zhang, Y. (2020). Ai explainability 360: An extensible toolkit for understanding data and machine learning models. Journal of Machine Learning Research, 21(130), 1-6.

Ateniese, G., Mancini, L. V., Spognardi, A., Villani, A., Vitali, D., & Felici, G. (2015). Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. International J of Security and Networks, 10(3), 137-150

Bandy, J. (2021). Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1-34.

Barber, D. (2012). Bayesian reasoning and machine learning. Cambridge University Press.

Barber, D. (2020). Private Machine Learning via Randomised Response. arXiv preprint arXiv:2001.04942.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943.

Bieker, F., Friedewald, M., Hansen, M., Obersteller, H., & Rost, M. (2016, September). A process for data protection impact assessment under the european general data protection regulation. In Annual Privacy Forum (pp. 21-37). Springer, Cham.

Bird, Sarah, et al. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft, Tech. Rep. MSR-TR-2020-32

Bloch, M., Barros, J., Rodrigues, M. R., & McLaughlin, S. W. (2008). Wireless information-theoretic security. IEEE Transactions on Information Theory, 54(6), 2515-2534.

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.

Budig, T., Herrmann, S., & Dietz, A. (2020). Trade-offs between Privacy-Preserving and Explainable Machine Learning in Healthcare.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., ... & Kurakin, A. (2019). On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705.

REFERENCES

Cave, S., Craig, C., Dihal, K., Dillon, S., Montgomery, J., Singler, B., and Taylor, L. (2018). Portrayals and Perceptions of AI and Why They Matter. The Royal Society, London. https://royalsociety.org/~/media/policy/projects/ai-narratives/AI-narratives-workshop-findings.pdf

Cave, S., Coughlan, K., & Dihal, K. (2019, January). " Scary Robots" Examining Public Responses to AI. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 331-337).

CDEI. (2021). The need for effective AI assurance. Available at: https://cdei.blog.gov.uk/2021/04/15/the-need-for-effective-ai-assurance/

CIPL. (2021). CIPL Recommendations on Adopting a Risk-Based Approach to Regulating Artificial Intelligence in the EU.

CMA (Competition and Markets Authority). (2021). Algorithms: How they can reduce competition and harm consumers. Available at : https://www.gov.uk/government/publications/algorithms-how-they-can-reduce-competition-and-harm-consumers

Cohen, J., Rosenfeld, E., & Kolter, Z. (2019, May). Certified adversarial robustness via randomized smoothing. In International Conference on Machine Learning (pp. 1310-1320). PMLR.

Cohen, S. N., Snow, D., & Szpruch, L. (2021). Black-box model risk in finance. Available at SSRN 3782412.

Cornelli, G., Doerr, S., Gambacorta, L., & Merrouche, O. (2020). Inside the regulatory sandbox: effects on fintech funding.

Dafoe, A. (2018). AI governance: a research agenda. Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK.

DataEthics (2020). Sandbox For Responsible Artificial Intelligence. https://dataethics.eu/sandbox-for-responsible-artificial-intelligence/

DataKind (2021). Harnessing the power of data science in the service of humanity. https://www.datakind.org/

Department of Industry, Science, Energy and Resources. (2018). AI Ethics Framework. https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework

Dong, Y., Fu, Q. A., Yang, X., Pang, T., Su, H., Xiao, Z., & Zhu, J. (2020). Benchmarking adversarial robustness on image classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 321-331)

Doteveryone (2020). Five years fighting for better tech, for everyone. https://doteveryone.org.uk/about/

Dwork, C., & Naor, M. (2010). On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. Journal of Privacy and Confidentiality, 2(1).

European Commission (February 2020). White Paper on Artificial Intelligence: A European approach to excellence and trust. Available at: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

European Commission. (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence

Fast, E., & Horvitz, E. (2017, February). Long-term trends in the public perception of artificial intelligence. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 31, No. 1).

REFERENCES

FAT/ML (2016) ‘Principles for Accountable Algorithms and a Social Impact Statement for Algorithms’. https://www.fatml.org/resources/principles-for-accountable-algorithms

Fish, B., Kun, J., & Lelkes, Á. D. (2016, June). A confidence-based approach for balancing fairness and accuracy. In Proceedings of the 2016 SIAM International Conference on Data Mining (pp. 144-152). Society for Industrial and Applied Mathematics.

Fredrikson, M., Jha, S., & Ristenpart, T. (2015, October). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (pp. 1322-1333).

Future of Life Institute (2021). Artificial Intelligence. Available at: https://futureoflife.org/artificial-intelligence/.

Ganju, K., Wang, Q., Yang, W., Gunter, C. A., & Borisov, N. (2018, January). Property inference attacks on fully connected neural networks using permutation invariant representations. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (pp. 619-633).

Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., ... & Holzinger, A. (2018, August). Explainable ai: the new 42?. In International cross-domain conference for machine learning and knowledge extraction (pp. 295-303). Springer, Cham.

Grant, T., & Wischik, D. (2020). Show us the data: Privacy, explainability, and why the law can’t have both. George Washington Law Review, 88 (6), 1350-1420. https://doi.org/10.17863/CAM.58412

Hall, P. (2019). An introduction to machine learning interpretability. O’Reilly Media, Incorporated.

Hitaj, B., Ateniese, G., & Perez-Cruz, F. (2017, October). Deep models under the GAN: information leakage from collaborative deep learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 603-618).

ICO (2020). Guidance on AI auditing framework: Draft guidance for consultation. https://ico.org.uk/media/about-theico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

ICO (2021a). Guidance on AI and data protection. https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/guidance-on-artificial-intelligence-and-data-protection/

ICO (2021b). Regulatory Sandbox. https://ico.org.uk/sandbox

ICO-Turing (2020). Explaining decisions made with AI Information. Information Commissioner’s Office & The Alan Turing Institute.

IPSOS Mori and Royal Society. 2017. Public views of Machine Learning: Findings from public research and engagement conducted on behalf of the Royal Society. IPSOS Mori Social Research Institute, London.

Johnson, D. G., & Verdicchio, M. (2017). Reframing AI discourse. Minds and Machines, 27(4), 575-590.

Kazim, E., & Koshiyama, A. (2020a). The Interrelation Between Data and AI Ethics in the Context of Impact Assessments. Available at SSRN.

Kazim, E., & Koshiyama, A. (2020b). A High-Level Overview of AI Ethics. Available at SSRN

Kazim, E. , Kerrigan, C. & Koshiyama, A. (2021a). EU Proposed AI Legal Framework. Available at SSRN 3846898 .

Kazim, E., Barnett, J., & Koshiyama, A. (2021b). Comments on CMA Report on How Algorithms May Reduce Competition and Harm Customers. Available at SSRN 3785647.

REFERENCES

Keskar, N. S., & Socher, R. (2017). Improving generalization performance by switching from adam to sgd. arXiv preprint arXiv:1712.07628.

Knowles, B., & Richards, J. T. (2021, March). The Sanction of Authority: Promoting Public Trust in AI. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 262-271).

Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., ... & Lomas, E. (2021). Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms.

Koshiyama, A., Firoozye, N., & Treleaven, P. (2020). Algorithms in Future Capital Markets. Available at SSRN 3527511.

Kouvaros, P., Lomuscio, A., Pirovano, E., & Punchihewa, H. (2019, May). Formal verification of open multi-agent systems. In Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (pp. 179-187).

Krafft, P. M., Young, M., Katell, M., Huang, K., & Bugingo, G. (2020, February). Defining AI in policy versus practice. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 72-78).

Larsen, J., & Hansen, L. K. (1994, September). Generalization performance of regularized neural network models. In Proceedings of IEEE Workshop on Neural Networks for Signal Processing (pp. 42-51). IEEE.

Le, L., Patterson, A., & White, M. (2018). Supervised autoencoders: Improving generalization performance with unsupervised regularizers. Advances in neural information processing systems, 31, 107-117.

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. Big Data & Society, 5(1), 2053951718756684.

Leslie, D. (2019). 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. https://doi.org/10.5281/zenodo.3240529

Lomas, E. (2010). Information governance: information security and access within a UK context. Records Management Journal.

Lomas, E. (2020). Information Governance and Cybersecurity: Framework for Securing and Managing Information Effectively and Ethically. Cybersecurity for Information Professionals: Concepts and Applications.

London, A. J. (2019). Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Center Report, 49(1), 15-21.

Longo, L., Goebel, R., Lecue, F., Kieseberg, P., & Holzinger, A. (2020, August). Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 1-16). Springer, Cham.

Lütge, C. (2020). White Paper on AI Ethics and Governance “Building a Connected, Intelligent and Ethical World”. TUM School of Governance.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635.

Melis, L., Song, C., De Cristofaro, E., & Shmatikov, V. (2019, May). Exploiting unintended feature leakage in collaborative learning. In 2019 IEEE Symposium on Security and Privacy (SP) (pp. 691-706). IEEE.

Meyer, M. (2014). Continuous integration and its tools. IEEE software, 31(3), 14-16.

Molnar, C. (2019). Interpretable machine learning. Lulu.com.

REFERENCES

Mueller, H. and Ostmann, F. (2020). AI transparency in financial services – why, what, who and when?. Financial Conduct Authority. Available at: https://www.fca.org.uk/insight/ai-transparency-financial-services-why-what-who-and-when

Nachev, P., Kennard, C., & Husain, M. (2008). Functional role of the supplementary and pre-supplementary motor areas. Nature Reviews Neuroscience, 9(11), 856-869.

Nicolae, M. I., Sinn, M., Minh, T. N., Rawat, A., Wistuba, M., Zantedeschi, V., ... & Edwards, B. (2018). Adversarial Robustness Toolbox v0. 2.2.

NIST (2020). Ongoing Face Recognition Vendor Test (FRVT). National Institute of Standards and Technology. Available at: https://pages.nist.gov/frvt/reports/11/frvt_11_report.pdf

OECD. (2021). OECD AI Policy Observatory. OECD.Ai. https://www.oecd.ai/

Orekondy, T., Schiele, B., & Fritz, M. (2019). Knockoff nets: Stealing functionality of black-box models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4954-4963).

Pagallo, U., Aurucci, P., Casanovas, P., Chatila, R., Chazerand, P., Dignum, V., ... & Valcke, P. (2019). AI4People-On Good AI Governance: 14 Priority Actions, a SMART Model of Governance, and a Regulatory Toolbox.

Partnership on AI. (2018). Tenets - The Partnership on AI. Available at: https://www.partnershiponai.org/tenets/

Qin, C., O'Donoghue, B., Bunel, R., Stanforth, R., Gowal, S., Uesato, J., ... & Kohli, P. (2019). Verification of non-linear specifications for neural networks. arXiv preprint arXiv:1902.09592.

Reisman, D., Schultz, J., Crawford, K., and Whittaker, M. (2018) Algorithmic Impact Assessments: A practical framework for public agency accountability. New York: AI Now Institute

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

Rodrigues, M., & Darwazeh, I. (2003, September). A spectrally efficient frequency division multiplexing based communications system. In Proc. 8th Int. OFDM Workshop (pp. 48-49).

Royal Society. (2020). Communicating AI: the role of researchers in enabling public conversations about AI. Available at: https://royalsociety.org/-/media/policy/projects/ai-and-society/how-we-talk-about-ai-and-why-it-matters-workshop-notes.pdf?la=en-GB&hash=E79120CDC567BEC262F33E0320159F93

RSA. (2018). Artificial intelligence: Real public engagement. London: RSA.

Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

Stanford HAI (2020). A Hub for Policy Impact. Available at: https://hai.stanford.edu/policy.

Stone, P. et al. (2016). Artificial intelligence and life in 2030. One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel

Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. Government Information Quarterly, 36(2), 368-383.

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Gonzalez Zelaya, C., and Van Moorsel, A.. (2020). The relationship between trust in AI and trustworthy machine learning technologies.InProceedings of the 2020Conference on Fairness, Accountability, and Transparency.272–283.

REFERENCES

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction apis. In 25th USENIX Security Symposium (USENIX Security 16) (pp. 601-618).

Treleaven, P., & Batrinca, B. (2017). Algorithmic regulation: automating financial compliance monitoring and regulation using AI and blockchain. Journal of Financial Transformation, 45, 14-21.

Treleaven, P., Barnett, J., & Koshiyama, A. (2019). Algorithms: law and regulation. Computer, 52(2), 32-40.

Ugwudike, P. (2020). Digital prediction technologies in the justice system: The implications of a 'race-neutral' agenda. Theoretical Criminology, 24(3), 482-501. https://doi.org/10.1177/1362480619896006

UK-CDEI (2020). Review into bias in algorithmic decision-making. Centre for Data Ethics and Innovation. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/957259/Review_into _bias_in_algorithmic_decision-making.pdf

UK Committee on Standards in Public Life (2020). Artificial Intelligence and Public Standards: report. Lord Evans of Weardale KCB DL. Available at: https://www.gov.uk/government/publications/artificial-intelligence-and-public-standards-report

UK-CQC (2020). Using machine learning in diagnostic services. UK's Care Quality Commission. Available at: https://www.cqc.org.uk/sites/default/files/20200324%20CQC%20sandbox%20report_machine%20learning%20in%20diagnostic%20services.pdf

US-DOD (2020). DOD Adopts Ethical Principles for Artificial Intelligence. Available at: https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificialintelligence/

Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. In 2018 ieee/acm international workshop on software fairness (fairware) (pp. 1-7). IEEE.
Wang, P. (2008, March). What Do You Mean by "AI" ?. In AGI (Vol. 171, pp. 362-373).

Weber Shandwick & KRC Research. (2016). AI-Ready or Not: Artificial Intelligence Here We Come!.

Weng, Y.H. (2015). Japan's robot policy and the special zone for regulating next generation robots. Tech and Law Center, Milan. Accessed at: http://techandlaw.net/japans-robot-policy-and-the-special-zone-for-regulating-next-generation-robots/

Xie, C., Wu, Y., Maaten, L. V. D., Yuille, A. L., & He, K. (2019). Feature denoising for improving adversarial robustness. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 501-509).

Yilmaz, E., Craswell, N., Mitra, B., & Campos, D. (2020, July). On the reliability of test collections for evaluating systems of different types. In proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2101-2104).

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., & Gummadi, K. P. (2019). Fairness Constraints: A Flexible Approach for Fair Classification. J. Mach. Learn. Res., 20(75), 1-42.

APPENDIX. RELATED WORK

Theme 1 – Public trust

Here we review work and surveys that have been carried out to estimate public awareness, perception and trust of AI systems. We then talk about the role and importance that the AI language and narratives can play.

Public Awareness, Perception and Trust

There is a common understanding that AI development will be conditioned on its acceptance and trust by the wider public (Ribeiro et al, 2016; Toreini et al., 2020; Arnold et al., 2019). Whilst it is important that the public is aware of the risks of AI - as to encourage regulation and best practice - misplaced fears regarding AI could hamper the development of potential beneficial applications (Cave et al. 2019; Stone, P. et al. 2016). In recent years, several surveys and studies were conducted to understand how people perceive AI risks and benefits. Results can depend greatly on the region as well as application of AI. Furthermore, the field and associated concerns evolve rapidly, hence one should be aware of which year the study was done.

- **AI awareness:** A UK survey commissioned by the Royal Society in 2016 (IPSOS Mori and Royal Society, 2017) found that most participants were unfamiliar with the term “Machine Learning”, but they were familiar with at least one of its applications. In another UK survey published in 2019 (Cave et al. 2019), 85% of respondents had heard of “Artificial Intelligence” and 42% gave a plausible definition, whilst 25% mentioned robots. Although these surveys were conducted a few years apart, it is interesting to see the difference in awareness between the terms machine learning and artificial intelligence. In a more global survey including participants from the US, Canada, the UK, China and Brazil (Shandwick and KRC Research, 2016), 18% of respondents stated to know a lot about AI (31% in China, 17% in Canada and Brazil, 14% in the US and 10% in the UK), 48% a little and 34% nothing. Further to the knowledge of the terms or what they represent, only 32% of people in the UK are aware that AI is being used for decision-making in general, and only 14 and 9% are aware that it is being used in the workplace and criminal justice system (RSA, 2018).

- **AI perception:** In a paper from 2017 (Fast et al, 2017), the authors analysed the content of the New York Times over a 30 years period to understand how public perception of AI evolved over time. They found a sharp increase in AI discussions since 2009, with consistently more optimism than pessimism. However in recent years, concerns over the loss of control of AI, ethical aspects of AI, and the negative impact of AI on work have been growing. In the Royal Society UK survey (IPSOS Mori and Royal Society, 2017), the authors found an even split between people thinking that the benefits of AI outweigh the risk and the ones thinking the opposite. Participants identified risks more easily than benefits, with the main concerns being that this technology could (i) harm them or others, (ii) replace them, (iii) depersonalise them and their experiences and (iv) restrict them. The main identified benefits were that this technology : (i) has a lot of potential to benefit individuals and society, (ii) could save a lot of time; and, (iii) could give them better choices. In (Cave et al., 2019), UK participants were asked how they feel about different “extreme” positive or negative scenarios involving AI, such as Immortality or Uprising. They found that on average across these specific narratives, the level of concern was higher than the level of excitement. In (Weber Shandwick and KRC Research, 2016), more than 60% consumers from China and Brazil expect a positive impact of AI on society against 38%, 21% and 18% for the US, the UK and Canada respectively, showing regional differences. The primary impression of AI for most participants in this survey came from the media. On top of the concerns are the criminal use of AI technology, job losses, cyber-attacks, a loss of privacy and humans becoming lazy.

- **The level of trustworthiness also greatly depends on the task or application:** In (Lee, 2018), the author conducted an experiment measuring perception of human vs algorithmic decision making in various scenarios. For tasks requiring “mechanical skills” (e.g. work scheduling), participants perceived the process as equally fair and trustworthy whereas for tasks requiring “human skills” algorithmic decisions were perceived as less fair and trustworthy. In (IPSOS Mori and Royal Society, 2017) and (Shandwick and KRC Research, 2016), it is also shown how perceptions differ for different applications. As an other example, an RSA/YouGov 2018 survey on AI and automated decision systems (RSA, 2018) showed that UK respondents were not supportive of the use of AI in decision-making, with 60% opposing or strongly opposing the use of AI in automated decision systems in the workplace and in the criminal justice system.

Lexicon, Narratives & Public Engagement
Gaining public trust is not a one-path way. A lot of efforts are underway to develop a more trustworthy AI where accuracy is no longer the only metric of success, but where explainability, fairness, privacy and robustness play a growing role. Alongside, different organisations and institutions, more are exploring the role of AI governance (e.g. regulation and standards). This awareness and evolution from practitioners and regulators will be a key step to gaining public trust, but public engagement, as well as the way the field communicates to the public, will also be important. We focus on existing research and efforts in the latter.

- **Finding common definitions:** Hampering this communication is the lack of common vocabulary and agreed-upon definition in the field (Krafft et al., 2020; Johnson and Verdicchio, 2017). This starts with the definition of AI itself (Stone, P. et al. 2016, Russell & Norvig, 2002; Wang, 2008). (Russell & Norvig, 2002) define AI according to four categories: systems that think: (i) humanly, (ii) rationally; or act: (iii) humanly, (iv) rationally. (Wang, 2008) splits the notion of AI according to five different research goals: Structure-AI (imitating the brain), Behavior-AI (imitating human behavior), Capability-AI (problem and application focused), Function-AI (building the best map from inputs to outputs) and Principle-AI (identifying the fundamental principles enabling to reproduce human intelligence). (Krafft et al., 2020) explore the gap between the way AI researchers and policymakers define AI. AI researchers tend to use definitions that emphasize technical specifications whereas policymakers were more likely to compare AI to human thinking or behaviour. They suggest that AI researcher definitions are more closely related to current AI technologies whilst policymakers may have a more "futuristic" view of AI. They propose to use a common 'policy-facing' definition that (i) includes both currently deployed AI technologies and future applications, (ii) is accessible for non-expert audiences, and (iii) allows for policy implementation of reporting and oversight procedures.

- **Reframing the discourse:** (Johnson & Verdicchio, 2017) also discuss the lack of clarity around the term 'autonomy', which is widely used by AI researchers for various computational behaviours, and can be misinterpreted by non-experts as a process over which humans have no control, causing potential worry and fears and "hiding" the true involvement of humans in the process. They stress the importance of developing and using a language that avoids such confusion across researchers, industry, policy makers and the wider public; and propose a new ontology to frame AI discourse. On a similar theme, earlier in 2021, the AI Now Institute launched a call for contributions in order to think about a new AI lexicon, stating that terms like "fairness, transparency and accountability" might not carry the same meaning in Western and non-Western communities, hence proposing to rethink the AI ethical discourse (AI Now Institute, 2021).

- **AI Narratives:** Along with the terminology, the narratives around AI play an important role. The Leverhulme Centre for the Future of Intelligence and the Royal Society held a series of workshops between May 2017 and May 2018 on these subjects in the context of the AI narratives project (Cave et al, 2018). They found that functional and non-fictional narratives revolve around futuristic technologies that are far from becoming a reality, or a very small subset of contemporary research. Although these narratives can be helpful in some ways, they can create exaggerated expectations or fears that can have "significant consequences for AI research, funding, regulation and reception". They suggest the addition of new narratives to the existing ones and the creation of spaces for public debate.

- **Perceived lack of control:** In (Cave et al, 2019), nearly 62% of the survey's UK respondents didn't feel like they were able to influence the development of AI, either because of their older age, technological determinism or because their views are simply not solicited. Participants felt that big businesses were in control of AI development as well as researchers and the government. People expressed a disconnection or distrust in these institutions, and felt the government had little control to regulate AI in businesses.

- **Efforts to engage with the public**: The RSA and DeepMind launched a "Forum for Ethical AI" to try and embed the public voice in the AI debate (RSA, 2018). Several other organisations are advocating or working on facilitating public access to the technology, transparency and accountability such as Partnership on AI (Partnership on AI, 2018); the AI Now Institute (Reisman et al., 2018); the Fairness, Accountability, and Transparency in Machine Learning community (FAT/ML, 2016); DataKind (DataKind, 2021); Doteveryone (Doteveryone, 2020) (now subsumed under the Ada Lovelace Institute).

Theme 2 – Accountability & operationalisation

There is a need to develop stronger regulation and standards, as well as frameworks for accountability. This will have in itself strong impacts on public trust (Knowles & Richards, 2021), but will also help ensure that AI is used in reasonable ways and beneficial to society. Although there is little to no regulation in place specific to AI right now, calls for more accountability are multiplying. It is expected that the satisfaction of particular standards will soon become mandatory (e.g. certification, auditability). These will be either general or sector specific (Koshiyama et al., 2021). We first present the current regulation or guidelines for general AI from governments around the world before exposing some related or sector-specific regulation already in place.

● **Existing regulation and standards:**
There are no current laws that specifically regulate the use of AI. However, many laws are applicable to the use of AI, such as GDPR, and other privacy legislation (AI systems must protect privacy), equality and anti-discriminatory laws (AI systems must not discriminate), and rights to recourse (citizens have a legal right to an explanation and hence, where appropriate, systems must be explainable). Notwithstanding the current lack of explicit 'AI regulation', there is a high active debate with proposed legislations (ex. Federal algorithmic Accountability Act, 2019). The most substantive of these is the European Commission recently published a draft of the first ever legal framework focused solely on AI (European Commission, 2021; Kazim et al., 2021a). This draft legislation proposes a risk based approach to AI governance.

There are four tiers:

• Limited risk: concerns systems that do not pose a threat to the safety and livelihood of persons. Action in this context pertains to 'transparency obligations' i.e. users should be aware that they are interacting with a machine in order to make an informed decision (article 52). Here self-regulation and mechanisms of adhering to codes of practice are appropriate.

• Minimal risk: This concerns systems that do not pose a threat to the safety and livelihood of persons are considered. Here the right to opt out of the use of such technologies and transparency provisions (ex. ensuring users are aware they are interacting with a machine), suffice. No action is necessary in this context and it is envisioned that the vast majority of systems will fall into this category. We have grouped these together and will not treat them further as they are of least concern to our interests in this white paper.

• High-risk: Here a general criterion is not offered, instead examples of sectors and applications are given (expanded upon below and corresponding to Title III, Chapter 2, articles 9-15). We infer from the case studies that, similar to unacceptable risk, such systems pose a threat to the safety and livelihood of persons, however, in these cases there are benefits that can be derived and used to justify deployment through good governance/risk management. In such high-risk cases a number of (legal) requirements are stipulated in terms of justifying the use of these high-risk systems. Indeed article 9 asserts the need to establish a 'risk management system' that must be acted upon and maintained, including adequate documentation. It is suggested that this is a 'continuous iterative process run throughout the entire [high-risk system's] lifecycle'. Following this, articles 10-15 denote, in more detail, the conditions that have to be met for a system to be justified for use.

- *Data and data governance (article 10): Training, validation and testing data sets to ensure that they are of high quality data.*
- *Documentation (article 11, 12): Provide detailed documentation for third party assessment, including technical documentation and record-keeping i.e. period logging of standards specifications being met.*
- *Transparency for users (article 13): comprehensible information regarding contact details of provider, purpose, accuracy, security, data used, human-oversight measures and expected life-cycle of a system should be reported.*
- *Human-oversight (article 14): Must ensure high-level of human oversight in development and deployment, through appropriate interfaces. The overseers must be able to understand the capacities and limitations of a system, avoid automatically accepting recommendations of a system, and be able to intervene effectively. Decisions should be taken after at least two people have overseen the system.*
- *Accuracy, robustness and cybersecurity (article 15): such relevant metrics must be declared, including failsafe mechanism, mitigation strategies against vulnerabilities and for cybersecurity attacks.*

• Unacceptable risk: Here concern is with systems that pose a direct and clear threat to the safety, livelihoods and rights of people. The action for such systems is an outright ban. Three use cases are named, these are:

- *Social Scoring systems: in opposition to systems that have been used in China, inferring character judgements from social behaviour is banned. Cases where a person incurs traffic incidents or engages in other kinds of antisocial behaviour should have no bearing on other (public) services/benefits they may receive.*
- *Manipulation: two kinds of algorithmic manipulation are discussed, namely that of vulnerability and that of the subliminal. Regarding the former, the elderly, children, and those with disabilities are noted. Regarding the former, we read reference to the 'algorithmic nudging literature, which concerns a person being 'nudged' i.e. manipulated by an algorithm towards a particular end, such as voting for a political party/candidate or purchasing. A qualifier 'significant' is introduced, however, it is noteworthy that no sustained treatment of what constitutes significant/subliminal manipulation is given.*
- *Remote biometrics: the use of indiscriminate scanning and use of identifiable characteristics (c.f. Facial recognition, audio scanning, sentiment analysis in the public sphere, etc.) are banned. The qualifier 'remote' is used to indicate that individual and consensual use of such systems is fine i.e. logging in via fingerprint, face, voice, etc*

In addition to the above, there are various initiatives and guidelines from governments around the world. The OECD website (OECD,2021) keeps a database of national strategies and policy initiatives around the world. In Canada, the government developed an Algorithmic Impact Assessment that policy makers and other officials should use to assess and mitigate the risks associated with deploying an automated decision system. In the UK, the ICO published guidance on AI auditing framework (ICO,2020), explainability (ICO-Turing, 2020) and AI and data protection (ICO, 2021a). Australia published an AI Ethics Framework with a set of voluntary AI Ethics Principles to follow (Department of Industry, Science, Energy and Resources, 2018).

● **Related or sector specific regulations:** Despite the lack of general regulation specific to AI, there exist laws in Data protection regulations such as GDPR that are very relevant to AI. Indeed Data Protection Impact Assessments are heavily referenced in the AI impact assessment literature (Kazim & Koshiyama, 2020a). There are also some sector specific guidance existing. For instance the Department of Defence in the US laid out guidance for use of AI in defense (US-DOD, 2020), the UK's Financial Conduct Authority is very active in the standards debate for AI systems in financial services (Mueller and Ostmann, 2020), and the UK's Care Quality Commission for medical diagnostic services (UK-CQC, 2020). Other application specific standards are being developed such as in Recruitment (UK-CDEI, 2020) or Facial Recognition (NIST, 2020).

● **Non-technical vs Technical Governanc:**. AI governance can be split into two categories: non-technical governance vs technical governance. 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" (Koshiyama et al., 2021). On the other hand, technical governance "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)" (Koshiyama et al., 2021). According to the European Commission (European Commission, 2020, 2021), technical auditing will focus on the four verticals (aka. challenges) mentioned in this report: (i) Robustness and Performance, (ii) Bias and Discrimination, (iii) Explainability and Interpretability, (iv) Privacy. There is currently a lot of research being done on both subjects by universities, independent organizations or within the industry.

- **Non-technical governance research:** Researchers and non-governmental organizations are doing tremendous work to propose guidelines or frameworks for auditing and impact assessment. In the UK we can for instance cite UCL with this new Algorithm and Standards lab, the Ada Lovelace Institute (Ada Lovelace Institute, 2021), Alan Turing Institute (Leslie, 2019), the Centre for the Governance of AI at Oxford University (Dafoe, 2018). In the US some examples are the AI Now Institute (Reisman, 2018), Partnership On AI (Partnership on AI.2018), Stanford HAI (Stanford HAI, 2020), the Future of Life Institute (Future of Life Institute, 2021) or the Centre for Information Policy Leadership (CIPL,2021). Other examples are the Ethics of AI Lab at Toronto University, the Institute for Ethics in Artificial Intelligence in Munich (Lutge, 2020). There are also a number of cross-organisational research efforts exploring AI governance (Koshiyama et al., 2021, Brundage et al., 2020). This workshop also made clear that industries such as Cisco or Unilever are developing their own internal set of ethical practices for AI.

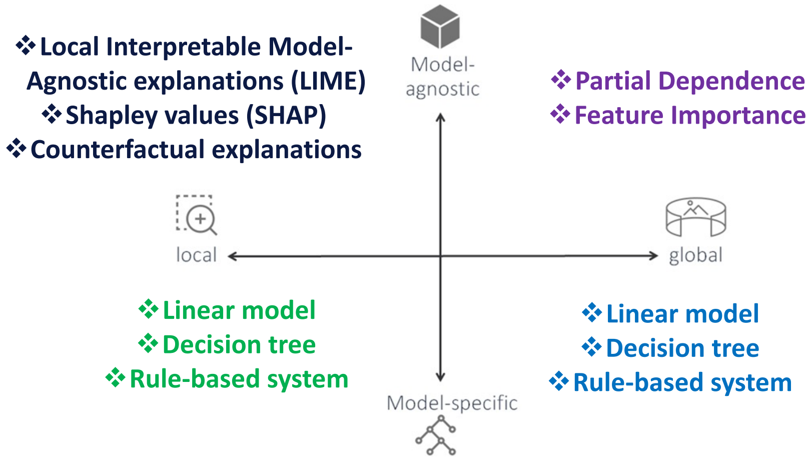
- **Technical governance research:** On the technical side, research to make algorithms themselves more fair, explainable, robust or private has been omnipresent in the past years. (Koshiyama et al., 2018) give a presentation of the different verticals and possible mitigation strategies. There is a vast amount of research done on each of these verticals and we will only give a quick overview of what the research focuses on, as well as some references for each. Much of the content is inspired or borrowed from (Koshiyama et al., 2018).

- **Fairness:** Numerous definitions and associated metrics have been proposed as well as mitigation strategies (Verma & Rubin, 2018; Mehrabi et al., 2019). We can differentiate between Individual and Group fairness: (i) Individual: the outcome for similar individuals should be the similar; and (ii) Group: some outcome measures should be equal across different groups (e.g. success rate, accuracy, etc.) (Koshiyama et al., 2018). Open-source toolkits based on current research are also being proposed to help practitioners assess and mitigate AI fairness such as Fairlearn by Microsoft (Bird et al., 2020) and AI Fairness 360 by IBM (Bellamy et al., 2018).

- **Explainability:** AI Explainability refers to the extent to which an AI system's inner workings and resulting decisions can be explained to humans. There are different methods used to explain algorithms depending on the model being used itself and the level at which we want to explain (population vs individual).
On Figure 4, we borrow a figure from (Koshiyama et al., 2018) that summarises these concepts. Here, model-specific explainability refers to models that are transparent and explainable by design such as linear or decision trees models, whilst model agnostic techniques are applied to opaque and complex models such as Deep Learning ones. Global refers to model explanation at "population" level whilst local at a more "individual" levels.

Some paper or book of references for Explainable AI are (Longo et al., 2020; Molnar, 2019; Hall, 2019). IBM also proposed an open source toolkit for explainability (Arya et al., 2020).

Figure 4: Types and levels of Algorithm Explainability (Koshiyama et al., 2018)



● **Performance and Robustness:** Performance and Robustness refer to the extent an algorithm is safe and secure, resistant to tempering or compromising of the training data. There are four criteria for performance and robustness: (i) Resilience to attack and security, (ii) Fallback plan and general safety, (iii) Accuracy and (iv) Reliability and Reproducibility. The first -Resilience to attack and security- states that an AI system should not be vulnerable to adversarial examples. Adversarial examples are inputs specially designed (by auditors or an attacker) to induce the model into making errors. Deep neural network are for instance particularly vulnerable to them (Dong et al., 2020), making adversarial robustness an important area of research. For related research, the reader can refer to (Carlini et al., 2019; Cohen et al.,2019; Xie et al.,2019). Again, IBM proposed a toolkit for robustness called Adversarial Robustness Toolbox v0. 2.2 (Nicolae et al., 2018). The second -Fallback plan and general safety- states that an AI system should have fallback plans in case of problems that are appropriate for the associated level of risks they carry. There are also sometimes constraints that a certain algorithm should verify. Formal Verification is carried out to prove the correctness of systems. Some useful references are (Qin et al., 2019; Kouvaros et al.,2019). The third -Accuracy- states that an algorithm should be able to make correct judgments (e.g. correct classification, prediction, decision). This concept notably asks: How well does my model accuracy, trained on specific data, generalize ? This is closely link with the Expected Generalization Performance (Arlot and Celisse, 2010). Papers exploring those issues are (Larsen & Hansen, 1994; Keskar & Socher, 2017; Le et al.,2018). The fourth and last -Reliability and Reproducibility- states that an AI system should be reliable (i.e. work for a range of inputs and situations) and reproducible (the behaviour should be the same should the experiment be repeated). This is usually enforced by developers through continuous integration (Meyer, 2014).

● **Privacy:** An AI system should maintain privacy and data protection throughout its entire lifecycle. There are three components of an AI system that needs to maintain privacy: (i) the data, (ii) the algorithm, and (iii) the interactions between the two. As we have mentioned earlier, the data is already regulated in many countries with laws such as GDPR that include Data Protection Impact Assessments (Bieker et al., 2016). Regarding the algorithm component, a privacy attack of concern is when an attacker is able to infer the model parameters and build equivalent models from this. Researchers have assessed models by reproducing possible attacks (Ateniese et al., 2015; Tramèr et al., 2016; Orekondy et al., 2019). Finally, the last component relates to the interactions between the data and the algorithm. The key concern here is the ability for an attacker to infer data about individuals from the population or training dataset through interactions with the algorithm. Possible attacks explored by the research community to help assess algorithms include model inversion (Fredrikson et al., 2015), statistical disclosure (Dwork and Naor, 2010), inferring prototypical samples of the targeted training set (Hitaj et al., 2017), membership and property inference (Ganju et al., 2018; Melis et al., 2019).

● **Interactions and Trade-offs:** The four verticals presented above are often studied in siloes, but interactions between them have also been explored. The first interaction we consider here is Accuracy/Robustness vs Explainability explored for instance in (London, 2019; Koshiyama et al., 2020). In general, more complex models are thought of as more accurate but more opaque than models such as decision trees or linear regression, although there are some exceptions (Koshiyama et al., 2021). The second is Fairness vs Accuracy/Robustness. Indeed, reducing biases can affect the overall accuracy (Fish et al., 2016; Zafar et al., 2019). The third interaction is Explainability vs Privacy (Budig et al., 2020; Grant & Wischik, 2020). The more a model is interpretable, the more difficult it is to maintain privacy. However, (Adriano et al., 2021) argue that using feature importance to explain an algorithm can aid with data minimization, hence making both explainability and privacy aligned. The last one is Fairness vs Explainability. In general, a more explainable model will help unveil biases and fairness concerns (Koshiyama et al., 2021).

Theme 3 – Experimentation

Participants to the workshops were unanimous in the fact that there was a need to build regulation and standards based on empirical evidence. Notably, it is important for industry to be able to share and get guidance on AI best practice. Note, as will be shown below, the need for experimentation comes from the lack of experimentation thus far (CMA, 2021; Kazim et al, 2021b).

● **Sandbox:** One way of doing this is through a regulatory sandbox. Indeed the proposed EU regulation on AI (European Commission, 2021) argues for regulatory sandboxes to ensure small enterprises can innovate in otherwise stringent compliance conditions. Regulatory sandboxes are quite common in the financial sector, with the first launched by the UK Financial Conduct Authority (FCA) in November 2015. Since then over 50 countries have adopted their own (Cornelli et al., 2020). Touching more closely on the AI sector, the ICO in the UK also created a Regulatory Sandbox to support organisations in the innovative and safe development of products and services which utilise personal data (ICO, 2021b). Inspired by this example, the Norwegian Data Protection Agency also created a Sandbox For Responsible Artificial Intelligence (DataEthics, 2020) which provides an environment for companies to develop responsible AI. (Pagallo et al., 2019) also advocate for “lawfully de-regulated special zones represent the legal basis upon which to collect empirical data and sufficient knowledge to make rational decisions for a number of critical issues”. The Japanese government set up such deregulated zones called Tokku to experiment with specific AI systems and help draw regulations (Pagallo et al., 2019; Weng, 2015). This includes road traffic laws (at Fukuoka in 2003, Sagami 2013), radio law (Kansai 2005), data protection (Kyoto 2008), and safety governance and tax regulation (Tsukuba 2011).

● **Pilot projects in AI assurance:** In the UK, the Centre for Data Ethics and Innovation launched a programme of work to understand the current maturity of assurance tools in industry, with notably four pilot audits conducted by a UCL team in collaboration with industry players (CDEI, 2021). These are looking to “address the gap of empirically informed discussions on standards and regulation in AI assurance”, and help identify what challenges might be encountered in practice and in different sectors. In (Sun & Medaglia, 2019), the authors analyse a practical case of adoption of the AI system IBM Watson in public healthcare to help understand the perceived challenges around such AI adoption. The empirical work ends with a set of guidelines for the governance of AI adoption in the public sector. In (Bandy, 2021), the author survey’s ‘audits’ and, on the spectrum of theoretical to empirically informed, is closer to the empirically informed side - however the author uses ‘audit’ very loosely to include all and any ‘review’ of a system. Nonetheless it is a useful recourse.

Although we have not found much evidence of such empirical work being conducted (perhaps because those who have conducted empirical work are reluctant to publish it), big industry players are developing their own set of practices, and enabling the sharing of information through tools such as sandboxes or conducting empirical pilot work in collaboration with the industry will be a priority in future years to ensure appropriate AI governance.