Published version citation: J. B. Peckham, "An AI Harms and Governance Framework for Trustworthy AI," in Computer, vol. 57, no. 3, pp. 59-68, March 2024, doi: 10.1109/MC.2024.3354040.

An AI harms and governance framework for Trustworthy AI

J. B. Peckham
Managing Director, Strategis Consulting Ltd.

***Abstract*—Many guidelines have been written for the development of trustworthy Artificial Intelligence (AI) and some frameworks proposed, but a common concern is the lack of precision in definitions that can make application difficult. I propose a novel governance and harms framework that seeks to provide more precision in the assessment and deployment of AI to meet trustworthiness objectives. Using a taxonomy of application types and associated potential harms, I show how four governance dimensions can be applied in any AI application to mitigate these harms. I highlight the technical limitations of achieving trustworthy AI solely through data selection and algorithm enhancement.**

***Keywords*—Trustworthy Artificial Intelligence, Ethics, Machine Learning.**

■ **INTRODUCTION** Over eighty bodies around the world have developed ethics guidelines for AI. Jobin in [1] provides a comprehensive survey of these guidelines and a summary of the ethical principles common to all is shown in **Table 1** adapted from [1].

There’s a good deal of overlap between the various reports and recommendations, with the importance of AI being for the common good standing out in most, alongside AI not harming people or undermining their human rights. Human rights are often the basis on which the idea of human autonomy is founded and relates to individual freedoms as well as the right to self-determination.

In recognition of the harms that can result from the use of some AI applications and negative user reaction, some groups have focussed on how to make them “trustworthy”. The European Commission High Level Expert Group (HLEG) is one example, producing specific guidelines for “Trustworthy AI” [2].

When it comes to preventing societal and individual harm from AI systems, although a laudable aim, these reports fail to spell out these harms adequately. What does it mean for systems to be safe and secure, or technically robust, given the nature of AI algorithms?

Perhaps one exception to this is the Centre for European Policy Studies (CEPS) report *Artificial Intelligence: Ethics, Governance and Policy Challenges* [3]. While containing many of the same terms as other ethics guidelines such as “non-maleficence” (do no harm), protecting human integrity, security and privacy, the report is a little more specific about what these might mean for AI applications and, usefully, lists some problematic use cases and no goers. One prohibited use case is autonomous weapons, while “problematic” examples are “predictive policing, social credit scores, facial recognition and conversational bots.”

What emerges from all these reports is that there's likely to be a tension between the potential benefits from using AI systems and the impact on individuals and society.

Table 1. Summary of key ethical issues and how they are described in various AI guidelines, listed in order of frequency of occurrence in 84 documents (adapted from Jobin, 2019).

ETHICAL ISSUE	DESCRIPTION
Transparency	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing.
Justice and fairness	Justice, fairness, consistency, inclusion, equality, equity, (non-) bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	Responsibility, accountability, liability, acting with integrity
Privacy	Privacy, personal or private information
Beneficence	Benefits, beneficence, well-being, peace, social good, common good
Freedom and autonomy	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	Trust
Sustainability	Sustainability, environment (nature), energy, resources (energy)
Dignity	Dignity
Solidarity	Solidarity, social security, cohesion

Whittlestone and her colleagues [4] suggest that while high- level principles in AI ethics are important, "they may not be enough to ensure society can reap the benefits and mitigate the risks of new technologies". The authors cite the example of bio ethics that similarly started with high-level principles, but in practice failed to deliver. They propose that the tensions between AI benefits and their negative influence should become the focus for AI ethical evaluations.

An example of tension cited by Whittlestone et al. surrounds a statement in the UK House of Lords AI Committee report [4, p. 197] that "it is not acceptable to deploy any artificial intelligence system which could have a substantial impact on an individual's life, unless it can generate a full and satisfactory explanation for the decision that it will take." Whittlestone et al., [4, p. 196] suggest that this statement "masks an important tension between using algorithms for social benefit (beneficence) and ensuring those algorithms are fully intelligible to humans (explicability)."

Many applications of AI are already in use today, such as assisting medical diagnosis and risk assessment, whose decisions, currently, cannot be explained, challenging how such applications might be deemed ethical and trustworthy.

Often the tensions are economic and the desire for efficiency can trump human rights. As the authors of *IEEE Ethically Aligned Design* report [5] suggest, "honoring holistic definitions of societal prosperity is essential versus pursuing one-dimensional goals of increased productivity or gross domestic product (GDP)."

What then, is the way forward for developing trustworthy AI applications, given these well-meaning but ill-defined high-level principles and the tensions that can occur? Can the challenge be met purely by better data selection for training and the development of explainable algorithms?

In this paper I will first outline what it means for an artifact, such as an AI application, to be trustworthy. I then define a set of governance requirements that can be used to address the ethical concerns in Table 1 and highlight the inherent limitations of implementing trustworthy AI, solely through algorithm development and better data selection.

I show how ethical concerns can be mapped into a taxonomy of AI application types and specific associated harms, providing a more actionable framework than the high-level ethical concerns. In the final section I propose a framework that sets the four governance factors orthogonal to the taxonomy of AI applications and harms, allowing governance requirements to be determined for each use case, even where multiple harms are identified.

DEFINING TRUST

The driver for trustworthy AI systems is the belief that their development, deployment and use will accelerate when society trusts them, leading to greater economic prosperity and human flourishing.

Trust is a fundamental component of societies but is based on trust between people rather than trust in artifacts. We may feel confident driving over a bridge because ultimately, we are putting our trust in the designers, builders and the regulators that it has been designed and built correctly and that materials have been checked along the way. This trust can be broken down when accidents occur because someone has not done their job properly, has cut corners or failed to heed warnings of structural deficiencies or erosion.

Trust in the context of AI is trust in the designers, companies and deployers of an application, that the application will do the job for which it has been created, correctly, reliably – that is produce the same result time and time again, without prejudice and will not do me any harm and that my experience of it will be beneficial. Our use of such applications will therefore entail a willingness to take risk – to trust the actors behind the design and deployment. Trust may be built over time as a result of repetitive use of an artifact where we build confidence that it does the job it was designed to do without harming us.

GOVERNANCE PRINCIPLES FOR TRUSTWORTHY AI

The IEEE P7001 Standard on Transparency for Autonomous Systems [8] approaches trust as a governance issue by creating five levels of transparency requirements for each of five different stakeholders. These are used to determine the level of transparency required (level 0-5) and the compliance for each stakeholder of any autonomous application. Explainability is regarded as a subset of transparency and relates to the extent to which information is accessible to non-experts.

In a similar way, other ethical concerns listed in Table 1, such as accountability and justice (also termed fairness) can be used to create a set of governance principles, each working together to provide an holistic framework to promote trust.

Figure 1 provides a definition and the implementation possibilities for each of the governance principles. Given the inherent difficulties of explaining the actions of stochastic AI algorithms, explainability in this paper addresses to what extent an AI algorithm can explain its actions or output and is the place to capture the risks associated with any unexplainable output. This is particularly relevant in applications that can cause harm because their output cannot be explained. On the flip side, recent developments in Generative AI, especially where natural language is output, are potentially harmful where they give the impression that the output can be explained to a non-expert user. Whether information is understandable by a non-expert is not the same as whether an algorithm can explain its output. In the case of Generative AI, even experts seem to be unable to explain so called "hallucinations" although it ought to be self-evident that a stochastic process could produce unexplainable output.

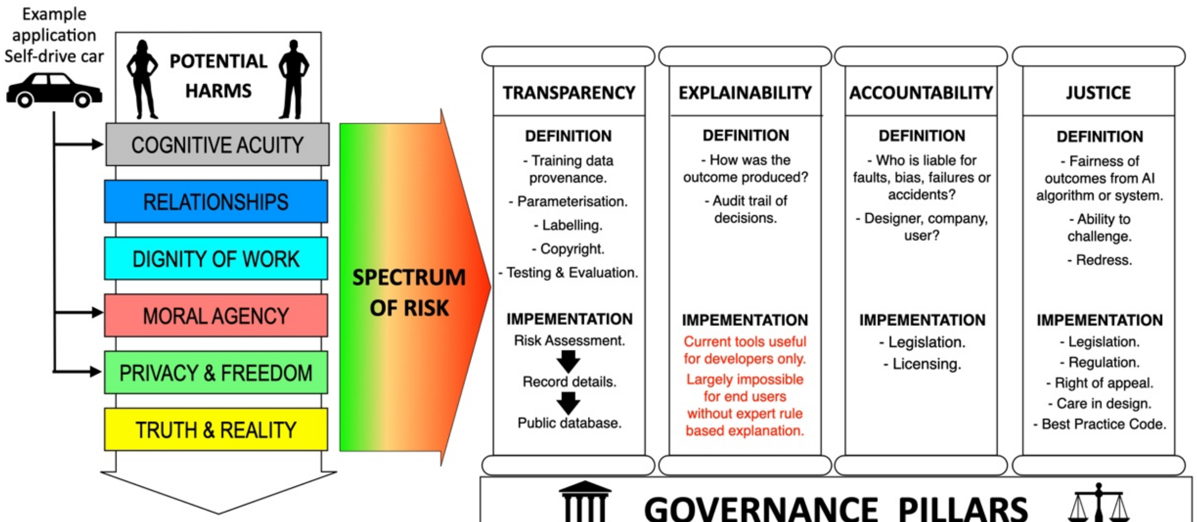

Figure 1. Mapping ethical principles into governance issues for AI applications and how they are implemented. Potential harms to humanity are on a spectrum of risk and are set orthogonal to these governance pillars. An application may have more than one type of harm.

These governance pillars are not separated from the other ethical concerns but become part of the framework for implementing measures that will promote trust. In the framework proposed here, the ethical concerns in Table 1 are mapped into an orthogonal taxonomy of AI applications and specific harms to users (Table 2) that exist on a spectrum of risk (see Figure 1). In a similar way, the application of IEEE P7001 [8] requires an Ethical Risk Assessment to identify the degree of transparency required for a particular application.

Transparency and accountability have straight forward ways of implementation, even if contentious in the case of accountability. In the other two areas of governance, explainability and justice, there are attempts to try and solve these challenges algorithmically and through better data selection and labelling for training. However, as we shall see, the nature of AI algorithms and training data make it difficult if not impossible to tackle these ethical issues automatically. Significance human intervention through governance will therefore be required to ensure AI trustworthiness in the sense that most policy makers intend.

I will now explore each of these governance areas in more detail before showing how they can be used alongside an orthogonal list of application harms to mitigate risk.

Transparency

Transparency is used by policy makers in a variety of ways from a simple data base recording the provenance of data sets, ownership and responsibility for algorithmic tools and impact assessments [9], to an ability for the output of the algorithm to be explained or its decision process uncovered [10].

The monitoring and use of data to continue model improvement in speech recognition for products like Amazon's Alexa, has caused public concern over data privacy. Whilst the human inspection and labelling of such data is enormously helpful to the developers and improves accuracy, the trade-off is the loss of privacy.

This clearly raises tensions between improving model performance and respecting users' privacy. The use of training data, both to initiate the model, and subsequent use of data to improve the model thus becomes a governance issue that cannot be resolved technically but is itself an ethical issue that will require regulation, perhaps in ways similar to Europe's General Data Protection Regulation (GDPR). Were such data to be unavailable by law without explicit informed consent, then further improvement in such algorithms would be slow, expensive and potentially limited by a lack of real-world data.

A requirement to register such details about an application would provide transparency to users and regulators and is one part of the process in protecting privacy. The accountability and justice component of governance may require the implementation of privacy laws to protect users' privacy where it is declared that data harvesting is carried out, for example in social media platforms. Where such laws exist, a transparency audit will reveal breaches (provided it is a regulatory requirement to file records).

Explainability

Explaining the reason behind an output from an AI algorithm, rather than the details of the mathematical process involved, is one of the most challenging aspects of AI development. Although there are attempts being made, it is unlikely to be solved technically because of the nature of the algorithms themselves.

Most advanced AI algorithms are better referred to as Machine Learning (ML) and although there are many variants, the most advanced tend to use Artificial Neural Networks (ANN) often multi-layered (so called deep learning). ANN based ML is a stochastic process that in simple terms produces a likelihood or probability of previously unseen input data matching the parameters of the training set. How the probability derived is unseen and untraceable, the so-called black box problem.

Human reasoning involves deduction, induction and abduction processes [11] and these are used by doctors in medical diagnosis or lawyers determining a case. AI algorithms are missing one crucial aspect, abduction, a process that no one yet has a theory for, so we can't encode it. In that sense the I in AI is a misnomer, there is no intelligence at all.

The typical challenges that image classifiers face, even those that use more generalisable models, include things like background clutter, view point variation and so called interclass variation (for example chairs come in many shapes and sizes). Humans on the other hand have no difficulty, in these scenarios, of quickly and correctly identifying the object. We use experience and can disambiguate difficult images, perhaps by forming a hypothesis to help us explain unusual features or even using intuition or guesswork. Crucially, humans are able to explain why they came to the conclusion they did, an AI system can't.

Abduction can be creative, intuitive or revolutionary involving leaps of imagination. Medical diagnosis is a good example of where clinicians use abduction, and explanation may be a critical aspect of the deployment of AI systems in this area. Zicari et al [12] used a hybrid approach of a rule based expert system where skin lesion classifier outputs also generate textual explanations, complemented by heat maps localising the criteria in the original image.

An overview of the approaches and challenges of explainable AI can be found in [13] and [14]. Bansal et al

[15] however found that many methods used to attribute an explanation for a classifier's decision were unstable and could give different explanations.

Whilst such approaches might eventually provide some level of explanation of the output of a classifier, Zicari et al., in [12] point out that "the explanations given will still depend on human analysis."

A conclusion from this consideration is that in applications where people's lives are impacted by an AI systems output, such as an Automated Decision Support System (ADSS), there should always be a right to request human analysis or even an opt out option. This dimension of governance can thus be used to explicitly flag up a need for specific legislation in the Justice dimension, thus providing a holistic framework for trustworthy AI.

Accountability

Accountability for faults, failures or harms arising out of the use of AI systems is a matter of legislation, that will require regulators and politicians to act. Self-drive or autonomous vehicles provide an example. In the U.K., the Independent Law Commission has proposed a new system of legal accountability [16] where "the person in the driving seat would no longer be a driver but a "user in charge"." This person would no longer be liable for offences that arise from the driving task. Responsibility would pass to an "Authorised Self-Driving Entity (ASDE) and in an authorised fully self-driving vehicles, people would become passengers, putting the licenced operator in charge and responsible."

Accountability in another application type, digital assistants, illustrates the challenges arising from the use of Natural Language Processing (NLP) systems such as deployed in Alexa. In a widely reported story [17], a ten year old child asking Alexa for a "challenge to do" received the reply "Plug in a phone charger about halfway into a wall outlet, then touch a penny to the exposed prongs." Fortunately, the child didn't carry out the challenge, but what if she had and who would have been liable if she had been electrocuted? The nature of such NLP systems and the fact that they have no understanding of the answers that they give highlights why accountability becomes a much more complex matter because the outputs of such devices cannot be predicted and additional ethical questions need to be asked to determine what, if any, governance concerns should be addressed, as we shall see later in the orthogonal harms dimension of my framework.

Justice

One of the most publicised aspects of unfair or unjust AI systems relates to biased data sets used for training ML algorithms, giving rise to discrimination. Obermeyer, for example, has pointed out in [18] that black patients were discriminated against in an AI system widely used in US hospitals. This was due to using the money being spent on patients as a proxy for health needs. On average, black people consume less resources, thus biasing the data when using this parameter as a proxy for health needs. This example is an obvious case where it is clear that improvements in design could be made to find better indicators of health need that reduce, if not eliminate, discrimination.

A major challenge, however, in data sampling and removing bias lies in determining the criteria to be used. Who determines what an unbiased data set looks like, given that ML is simply modelling on historic data as a proxy for the reasoning and abduction that humans carry out? Humans themselves are biased and this bias is encapsulated in historic data, often used for training. Using algorithmic methods or human screening simply introduces another form of bias into the data. Historic patterns of crime, as an example, might simply be a starting point in human evaluation but if crime predominantly occurs amongst a particular socio-economic group, trying to balance the data set in some way will itself introduce another bias. What is regarded as social bias is a concept that is not fixed in any given society and can change over time. It also differs between different groups and ideologies.

The process of uncovering bias is an iterative one that requires reasoning, empirical evidence, abduction and value judgement. In the final analysis, justice needs to be served through a right of appeal to human adjudication, rather than relying on a black box.

Data bias can also occur from malicious bias. An independent data audit through a Distributed Ledger Technology (DLT) is proposed by Thiebes in [19] and has the benefit of at least identifying malicious bias or intent. This approach is already being trialled to secure the provenance of news and images. Safe.Press, a news certification service developed by Block Expert in France uses blockchain technology to combat fakenews. The Content Authority Initiative with partners including Adobe, Twitter and the BBC is another venture seeking to address the problem using technology developed by start-up company, TruePic. TruePic also uses blockchain technology but is developing a more scalable system using public/private keys based on a Public Key Infrastructure.

At best, just outcomes could be improved through careful design, data sampling and labelling, but it is naïve to assume that data bias can be eliminated and that a data set could ever be complete.

MAPPING ETHICAL ISSUES TO HARMS

The key ethical principles surrounding the use of AI systems in Table 1, with the exception of beneficence, reflect the potential for harm in ways other than purely physical. The challenge with such principles is that they lack specificity and are used to convey a sense of what AI

Table 2. A taxonomy of AI applications and some of the harms that can occur to people from the use AI applications and systems that seek to simulate human attributes and capabilities.

Taxonomy of AI applications	Harms to Humanity – Loss of
AI replaces cognitive skills Autonomous Decision Support Systems - where these are used to determine outcomes for people where their lives could be seriously impacted – e.g., in finance, medicine, the judiciary, self-drive vehicles.	Cognitive Acuity When AI learns and carries out skilled tasks that humans perform, replacing these tasks with automation leads to a loss of reasoning power, decision-making acuity and creativity.
AI simulates humanness and/or creates bonding Voice digital assistants, chat bots and humanoid robots - that simulate human characteristics such as speech, hearing - the ability to interact verbally or visually. - simulation of sentience through altered speech (e.g., showing emotion) and/or expression (e.g. sadness, surprise), - detection of human feelings by speech and facial expression analysis, - reaction to touch.	Relationships Over-engagement with digital assistants, robot toys, healthcare robots and the use of sex robots fosters personification of artefacts and the development of non-human relationships that alters our ability to maintain or form true relationships with other humans. Our children's emotional and social growth is stunted and their ability to empathize is diminished along with the emotional maturity needed in normal human relationships and social interactions. Personification of artefacts leads to feelings of ethical obligation and the desire to assign rights to personified artefacts, amounting to idolatry.
AI is used for surveillance of citizens, and personal data is exploited by companies Facial recognition, gait analysis. Collection of personal data - from online activity or use of sensors (e.g., from IoT). Recommender algorithms - to increase user engagement, filtering posts or feeds that are manipulative.	Freedom to choose and privacy This results from the state's surveillance of its citizens whether through facial recognition and other traits or the amassing of private data for running smart cities. Freedom and privacy are lost due to the even greater amassing and processing of personal data by Big Tech for profit without any real choice for consumers. The free product or service offering model is an abuse of power because consumers are seduced by Big Tech's offerings without informed consent to their data use, which in any event would be impractical.
AI is given moral autonomy Autonomous weapons. - life is at risk. Fully autonomous Decision Support Systems. - where the output has significant impact on a life.	Moral Agency When we allow an artifact to perform actions on our behalf that might have moral consequences, we effectively cede moral agency, a capacity that's uniquely human.
AI replaces manual and skilled work Computer systems or robots	Work The dignity of work is taken away as jobs are partially or completely replaced by AI and robots, except where the work is hazardous.
AI is used to create a virtual reality on its own or to augment the real world Immersive games, virtual engagement either social or professional.	Reality A loss of a sense of what's real through immersion in virtual and augmented reality, a loss of self-discipline, self-determination and control through addiction with a resulting loss of true community from isolation and virtual relationships.

systems should or should not do to users. Some, like Transparency, become the focus for ethical concerns and require a separate Ethical Risk Assessment that can lack specificity.

I propose a different approach using a taxonomy of AI applications against which a more specific set of harms to humanity are defined as shown in **Table 2** (adapted from [20]). These seek to be more descriptive of the harms that can be caused to users of each particular use case of an AI system in the taxonomy.

The potential for subjectivity in assigning an application to a particular risk category in this taxonomy might be seen as problematic, especially in cases where an application might fit into different categories. In this framework, each category of harm should be considered, initially in the Transparency pillar of governance, and the risk assessed. A self-driving vehicle therefore would be assigned to several potential harm categories, such as replacing cognitive skills, used for surveillance (sensor data monitored by government or a company) and given moral agency. Each represents different risks but each needs to be considered.

The reduction in cognitive skills where a person loses skill at driving, might be regarded as acceptable and requiring minimal legislation. The surveillance potential using vehicle sensor information would prompt consideration of the governance needed around security and privacy. The moral dilemmas surrounding safety (e.g. should the vehicle be programmed to prioritise the safety of its occupants over others) might lead legislators to determine that self-drive vehicles should never be fully autonomous on the grounds that assigning liability could be problematic. Used in this way, rather than creating bias, the taxonomy allows different risks to be assessed and appropriate governance to be developed.

By collecting different harms and risk levels under the Transparency pillar of governance, the tensions between different principles can also be revealed and made more explicit leading to greater transparency. Where these are traded off by a stakeholder, whether the company selling the device or government institutions deploying or legislating for their use, these can also be captured in the Transparency process and made a disclosure requirement.

The assessment of Transparency requirements in IEEE P7001 is required for each of a number of defined stakeholders such as end users, safety inspectors and lawyers. Applying such an approach to the Transparency pillar in this paper adds depth to the proposed framework and could also be the focus for assessing the risks for each stakeholder.

In most use cases, there is a spectrum of risk associated with the deployment and use of the application types described in the taxonomy (see Figure 1). The German Data Ethics Commission proposed a five-level risk-adapted regulatory approach to algorithmic systems [21]. At the lowest level are applications where there is negligible or no risk whilst the highest level represents applications with an untenable potential for harm and should be banned or partially banned. These risk levels can be captured in the framework proposed here through the Transparency pillar as each application and harm category are evaluated (see Figure 1).

HARMS VERSUS GOVERNANCE FRAMEWORK FOR TRUSTWORTHY AI

Mapping ethical issues into a taxonomy of specific harms and a set of orthogonal governance principles allows a much sharper focus for evaluating any particular application of AI. For each application it is possible to assess each of the potential harms and the level of risk through the governance process, starting with transparency and explainability. These assessments then inform the more legal aspects of governance providing more specificity on what legislation and compliance with standards might be required. The governance axis therefore becomes the main focus for implementation, based around the risk assessment of the harms in a particular use case in the taxonomy (Table 2). It also highlights the inherent limitations of algorithmic and data selection solutions to the explainability and justice dimensions of governance that will require human intervention and oversight in ensuring trustworthiness.

There are many uses of AI where we have freedom to choose and to exercise self-control, either by moderation or abstinence, whether they be digital assistants such as Google Home, Satnav systems, browsers or tools in our workplace. The challenge arises when we have no control or choice over whether we’re subjected to applications, such as facial recognition, emotion detection or an ADSS. This will be the case when they’re used by the state in the public sphere without our consent. The harms principles would at least alert stakeholders in the deployment of such systems whether freedom of choice should be built in as a governance principle, for example, a patient could be given the opportunity to opt out of the use of a medical diagnostics tool.

The debates continue about what applications should be totally banned because of the risk associated with them. The trade-off is usually economics and efficiency yet, as the authors in [5] noted, this shouldn’t be the only criteria. I would argue that humans should retain responsibility for decision-making in areas where the lives of others are affected. For governments and corporations that means that we should not entrust to an algorithm the responsibility to make risk judgements about parole, reoffending, children potentially at risk, visa applications and many more areas.

The risk is one of mission creep, the DSS becomes autonomous, at the very least there should be a legal right of appeal to a human where abductive reasoning can be used in assessing a person's case.

The development, deployment and use of AI algorithms and systems should therefore be constrained by the human evaluated risks associated with a particular AI use case and appropriate governance put in place to protect the user. Where applications pose inconsequential risk of violating the principles then governance requirements might be light. Where there is significant risk, governance would be much more specific, even to the point of legal banning of an application. The example of facial recognition technology illustrates how this could work out in practice in different use cases.

Image recognition algorithms used in face recognition have considerable limitations cited earlier. When used for security in a smart phone, it is used by choice and the consequences of unreliability might be regarded, if not as inconsequential, as minor - but we have a choice whether to buy such a phone or to use that method of security. Our personal data in the form of a model of our face remain private. We are balancing risks of misuse where valuable data might be stolen against convenience. Given the limitations of the technology, governance concerns might suggest that devices should contain a warning to highlight the risks, including performance data that might point up that accuracy with some ethnicities might not be as good as others. We could regard this as a transparency issue that might also limit the company's liability due to prior disclosure.

On the other hand, if freedom and privacy are to be respected along with a recognition of the limitations of the reliability of the technology, then facial recognition should not be used as a determiner of identity, and biometric data generally should not be collected, stored and used without consent by the individual. In public uses of facial recognition,

freedom and privacy are directly assailed and a ban could be mandated. Some cities in the US have already taken this step and banned their use in public places [22].

CONCLUSION

A taxonomy of AI applications and categories of harm have been described as a first step towards improving the specificity of the ethical criteria used to evaluate and develop trustworthy AI. I have proposed a set of governance principles or pillars that are orthogonal to this taxonomy that can be used to capture the risks and trade-offs from multiple harms, where they exist, for any given application of AI. These become the main focus for the implementation of trustworthy AI. More research is needed to test out this approach in a variety of use cases and the AI taxonomy and harms definitions may need refining and extending in the light of experience. The environmental sustainability dimension could certainly be added.

Using this framework, I showed the limitations to developing trustworthy AI from a purely technical perspective in respect of transparency, explainability, accountability and justice. These were due to inherent, and likely ongoing, limitations of AI algorithms.

The major harms identified lie on a spectrum of risk [21] and these should be evaluated for a particular use case from the perspective of all stakeholders, that could be captured in the Transparency pillar in the same way that IEEE 7001 does. This would lead to the implementation of appropriate governance in the four-dimensions cited that might include appropriate warnings or restrictions in the applications use. A color-coded labelling system could be used to highlight and explain any warnings or potential harms, rather like food labelling.

The proposed evaluation process is offered as a contribution to the research and debate around the creation of AI applications that can be trusted by professionals and the public at large, to do the job they were designed for, without causing harm. Along the way some hard decisions will be required to set appropriate governance in place. The Dutch child care scandal over the use of ML to perform fraud risk assessment [23] could have been averted had the governance measures that I propose been implemented in recognition of the potential for significant harm and injustice to be perpetrated to the child benefit claimants. Perhaps above all, practitioners and users need to question whether AI should be used in some applications at all.

■ REFERENCES

1. A. Jobin, M. Ienca, and E. Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019.
2. Independent High Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI*, European Commission, 8 April 2019.
3. A. Renda, Artificial Intelligence: Ethics, Governance and Policy Challenges, Brussels: Centre for European Policy Studies, 2019, pp. 56–57.
4. J. Whittlestone, R. Nyrup, A. Alexandrova and S. Cave, The Role and Limits of Principles in AI Ethics: Towards a Focus of Tensions, *Proceedings of the Association for the Advancement of Artificial Intelligence Conference,* January 2019, pp. 195–200.
5. IEEE Global Initiative on the Ethics of Autonomous and Intelligent Systems, *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, IEEE, 2019,
6. R. Schwartz, J. Dodge, N. A. Smith, O. Etzioni, Green AI, *Communications of the ACM*, December 2020, Vol. 63 No. 12, Pages 54-63.
7. A. Lacoste, A. Lucioni, V. Schmidt, T. Dandres, Quantifying the Carbon Emissions of Machine Learning, *ArXiv* abs/1910.09700 (2019): n. pag.
8. Winfield AFT, Booth S, Dennis LA, Egawa T, Hastie H, Jacobs N, Muttram RI, Olszewska JI, Rajabiyazdi F, Theodorou A, Underwood MA, Wortham RH and Watson E

IEEE P7001: A Proposed Standard on Transparency. In *Front. Robot.* AI 8:665729. 2021 doi: 10.3389/frobt.2021.665729.

9. Algorithmic Transparency Standard, UK Government, [Online]. Available: https://www.gov.uk/government/collections/algorithmal-transparency-standard
10. Top 10 Principles for Ethical AI, UNI Global, 2017.
11. E. Larson, *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do*, The Belknap Press of Harvard University Press: Cambridge, Massachusetts, 2021. (book)
12. R. Zicari et al, (2021). Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier, *Front. Hum. Dyn* 3:688152. doi: 10.3389/fhumd.2021.688152
13. S.T. Mueller, et al., Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI, *arXiv preprint arXiv:1902.01876,* 2019.
14. A. B. Arrieta, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020): 82-115.
15. N. Bansal, C. Agarwal and A. Nguyen, SAM: The Sensitivity of Attribution Methods to Hyperparameters, CVPR 2020. [Online]. Available: http://s.anhnguyen.me/sam_cvpr2020.pdf
16. Legal reforms to allow safe introduction of automated vehicles, *Law Commission*, U.K., 26 January 2022. [Online]. Available: https://www.lawcom.gov.uk/legal-reforms-to-allow-safe-introduction-of-automated-vehicles-announced/
17. Alexa tells 10-year-old girl to touch live plug with penny, *BBC News*, 28 December 2021. [Online]. Available: https://www.bbc.com/news/technology-59810383
18. Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, *Science*, vol. 366, no. 6464, pp. 447–453, Oct. 2019, doi: 10.1126/science.aax2342.https://www.ftc.gov/system/files/documents/public_events/1548288/privacycon-2020-ziad_obermeyer.pdf
19. S. Thiebes, S. Lins and A. Sunyaev, Trustworthy artificial intelligence. *Electron Markets* **31,** 447–464, 2021. [Online]. Available: https://doi.org/10.1007/s12525-020-00441-4
20. J. Peckham, *Masters or Slaves? AI and the Future of Humanity*, IVP: London, 2021, p. 187. (book)
21. Opinion of the Data Ethics Commission – Executive Summary. [Online]. Available: thttps://www.bmj.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.pdf?__blob=publicationFile&v=2
22. City of Oakland letter to members of the City Council and Members of the Public, 6 June 2019. [Online]. Available: https://www.eff.org/files/2019/11/12/ oaklandfr.pdf.
23. B. Sledam and A. Brenninkmeijer, The Dutch benefits scandal: a cautionary tale for algorithmic enforcement, *EU Law Enforcement*. [Online], Available: https://eulawenforcement.com/?p=7941

Jeremy Peckham is Managing Director of Strategis Consulting Ltd., Bewdley. DY12 1LD, U.K His research interests include AI, AI ethics and leadership. Peckham received his Bachelor of Science degree in Applied Science from the University of Brighton. He is a Fellow of the Royal Society for the Arts. jeremy@strategis-consulting.co.uk