**Review on Responsible AI for conservation**

*( Wearn, O. R., Freeman, R., & Jacoby, D. M. (2019). Responsible AI for conservation. Nature Machine)*

1. First of all, it is obvious that AI(Artificial Intelligence) and ML(Machine Learning) are used in many fields for protecting the nature:
   1. predicting the extinction risk of thousands of species;
   2. assessing the global footprint of fisheries;
   3. and identifying animals and humans in wildlife sensor data recorded in the field.

Commercial companies supports:

1. Microsoft’s AI for Earth
2. Google’s AI for Social Good
3. However, their( AI and ML) misuse could have severe real-world consequences for people and wildlife. The opaque nature of some ML algorithms means that the potential for unintended consequences may be high and this could have real-world consequences for people and wildlife:
4. it can be difficult to identify the implicit assumptions of an algorithm (for example, how much of the contextual background information it is using when identifying species in images).
5. it might be unclear when an algorithm is being asked to make predictions beyond the scope of the training data.
6. an algorithm might not be easily interrogated as to why it made a particular decision.
7. Better metrics are needed, since simple accuracy metrics are unlikely to provide a good indicator of success when an algorithm is transferred to new datasets.
8. Better ethical oversight of the use of AI in conservation is needed.
9. Two potential goals for the conservation and AI communities to tackle in the immediate term:
   1. the development of metrics to better allow conservationists to assess the usefulness of an algorithm;
   2. and the formulation of ethical guidelines for the responsible use of AI in conservation.

**Review on Five Ps: Leverage Zones Towards Responsible AI**

Nabavi, E., & Browne, C. (2022). Five Ps: Leverage Zones Towards Responsible AI. arXiv preprint

1. What is Five Ps?

problem, parameter, process, pathway, and purpose

1. What can Five Ps do for responsible AI?
   1. Problems identified in the Parameter zone are tractable (modifiable, mechanistic) characteristics of an AI system that are commonly targeted by AI developers to improve the responsibility of AI. They are typically smaller visible flaws that are usually addressed through engineering solutions such as tweaking algorithms and parameters. The effort to fix these is small, and changes in this zone are incremental and may have a negligible effect on the problem’s underlying structure or dynamics. They are important markers of the problem, but they are often symptomatic and not the root cause of the problem.
   2. Problems identified in the Process zone consider the wide range of interactions between the feedback elements of an AI system that drive the internal dynamics, including social and technical processes associated with how the AI is designed, built, and deployed. This might include activities that speed up development times, or actively responding to emerging trends in the data. Changes in this zone are likely to result in resolving issues as they emerge or amplifying the effect of assumptions.
   3. Problems identified in the Pathway zone consider the ways through which information flows, the rules are set, and the power is organized. For example, improving transparency of how algorithms are employed, the governance or legislation of their use, or putting the ownership of data back into the consumer’s hands. These changes are structural to the system that allows the AI to operate, and result in establishing new patterns of behavior and agency.
   4. Issues identified in the Purpose zone have the most potential to affect change in a system. These relate to the norms, values, goals, and worldviews of AI developers that are embodied in the system. It includes the underpinning paradigms based on which the system is imagined, and the ability to transform entirely and imagine new paradigms. Framing perceived problems in this zone serves to act as a compass to guide the developers to align with the fundamental purpose of the system.

Ghallab, M. (2019). Responsible AI: requirements and challenges. AI Perspectives, 1(1), 1-7

1. How to foster research and development efforts toward socially beneficial applications?

**AI for the social good**

For example, the AI for Global Good Summit of the ITU is concerned with encouraging R&D in AI to actively contribute to the 17 Sustainable Development Goals (SDGs) of the UN.

The challenges for fostering AI toward social good fit in two main categories: *incentives* and *integrative research*.

**Incentives**

more support is needed from international cooperation and public funding, which should bring significant and concentrated resources on key objectives.

**Integrative research**

Integrative research within AI is demanded for addressing heterogenous tasks, which are inherent to socially beneficial applications. They also require the involvement of non-academic contributors, social actors and stakeholders within investigations and developments. Integrative research is intrinsically difficult. It requires

a long-time span, due in particular to the overhead of collaborations and field tests.

1. how to take into account and mitigate the human and social risks of AI systems.

**Mitigating AI risks**

There are a few general categories of risks that are common to many applications. These are notably: *(i)* the *safety* of critical AI applications, *(ii)* the *security and privacy* for individual users, and *(iii)* the *social risks*.

**Safety critical AI applications**

AI techniques are frequently integrated within artifacts and systems endowed with sensory-motor capabilities and increasing levels of autonomy. Health, transportation, network management, surveillance and defense systems. *Verification and Validation* (V&V) methods to AI and their industrial deployment. It is essential to be able to accurately analyze and qualify the safety properties of components and systems using AI.

**Security and privacy for individual users**

An associated querying engine must interpret each request in its context and in relation to the user’s profile, which is constantly learned, refined and evolving.AI can provide insight about where research and education efforts should concentrate. AI mediated interactions raise social risks (covered in Social risks), as well as individual risks. Security of digital interactions, Confidentiality, privacy and use of personal data, Intelligibility and transparency.

**Social risks**

1.take into account the long term, including possible impacts on future generations.

2.social cohesion.

**Biases, Behavior manipulation, Democracy, Economy, Employment, Military systems**

The needs for responsible AI developments with respect to the social risks correspond in particular to political and legal measures and to international agreements. A proactive approach must rely on *social experiments*, and integrative research about social risks and mitigation measures. Here too, a change of paradigm is required to fund and develop joint investigations between AI and social scientists, to give a better understanding of AI to the former, and of social and economic mechanisms to the latter.

**The growing effectiveness of AI is simply commensurate with its social responsibility.**

**Review on Design Method for Responsible AI**

Main Question: Can AI artefacts be designed to be verifiably ethical?

• Systematic attempt to include values of ethical importance in

design

• Make values, their priorities and choices explicit, transparent

and ‘verifiable’

A value is a concept or aspect that a person or group of people deem to be important in life.

Decisions Matter!

Values-----interpretation--🡪norms-----concretization--🡪functionalities

**FROM VALUES TO NORMS:**

Norms are based on reasons for action

Norms are rules for action

Norms are not always obligations/prohibitions but can also be recommendation

**FROM NORMS TO FUNCTIONALITIES:**

Context-dependent

Requires more information

o Scope of norm

o Specification of goals

o Specification of means

Three areas: Ethically acceptable, Legally allowed, Socially accepted

Fairness: Equal resources and Equal opportunity

Implementation: deal with the black box(unseeable)

Governance: deal with the glass box ( be transparent and open), verify limits to action and decision, define the ethical borders, have monitor for governance

First, Value. Second, Norm. Third, Implementation. Last, Governance

Four areas for Implementation choices:

Collaboration: Shared awareness • Explanation • Real-time decision

Regulation: Formal ethical rules • Institutions • Offline reasoning

Algorithmic: Formal ethical rules • Ethical reasoning • Real-time reasoning • Learning ethics

Random: Trust

**VERIFICATION AND VALIDATION**

It is not only about the AI system but the whole system

o How is used

o Who is using

o Purpose

o Environment

•Use(rs)

•AI

•Governance

Zhu, L., Xu, X., Lu, Q., Governatori, G., & Whittle, J. (2022). AI and Ethics—Operationalizing

Responsible AI. In *Humanity Driven AI* (pp. 15-33). Springer, Cham.

<https://arxiv.org/pdf/2105.08867.pdf>

This chapter focused on how to design, develop, and validate AI technologies and systems responsibly (i.e., Responsible AI) so that we can adequately assure ethical and legal concerns, especially pertaining to human values. This chapter discussed the challenges related to operationalising ethical AI principles and presents an integrated view that covers high-level ethical AI principles, general notion of trust/trustworthiness and product/process support in the context of responsible AI, which helps improve both trust and trustworthiness of AI for a wider set of stakeholders.

1. discusses the challenges of existing works on ethical AI.

In general, the existing works on ethical AI could be classified as three large categories: High-level ethical principle frameworks; Ethical algorithms; Human values in software engineering and their operationalisation. and issues in current research work are identified as three types regarding operationalising ethical principles to achieve the ultimate trust from stakeholders, they are:

* Mixing Inherent Trustworthiness with Perceived Trust

Trust is a stakeholder’s subjective estimation of the trustworthiness of the AI system, which is based on a truster’s expected and preferred future behaviour of the AI system. Mixing the two in terms of identifying assurance mechanisms and presenting trustworthy evidence can overlook the additional and special mechanisms required to gain trust

* Operationalising Ethical Principles

one of the important factor is due to the relatively narrow attempt to op- erationalise human values and ethical principles into verifiable “product” trustwor- thiness (via mathematical guarantees) without systematically exploring a wider variety of mechanisms in development processes to improve both trustworthiness and trust.), and

* Unique Characteristics of AI

AI would be impossible to accurately and completely specify all the goals, undesirable side-effects and constraints (including ethical ones) at its finest level of detail.

1. The framework with an integrated view of three aspects of ethical AI

* The difference Between Trust and Trustworthiness in the Context of Ethical AI Principles (as Software Qualities & Governance Issues)
* How different product and process mechanisms can achieve trustworthiness for different ethical principles
* How different product and process evidence can be presented to different types of trusters to improve the accuracy of their subjective estimation so they match the inherent trustworthiness of the systems.

1. shares their experience and observations in a crop yield prediction project from the following aspects

* Human, social and environmental wellbeing
* Human-centred values
* Fairness
* Privacy and Security
* Transparency and Explainability
* Contestability
* Accountability

1. talks about the high-level ethical principles and their operationalization

             Interpretation of high-level ethical principles from:

* Autonomy
* Explainability
* Accountability

             Operationalising AI ethics from:

* Requirements
* Design
* Operations
* Governance

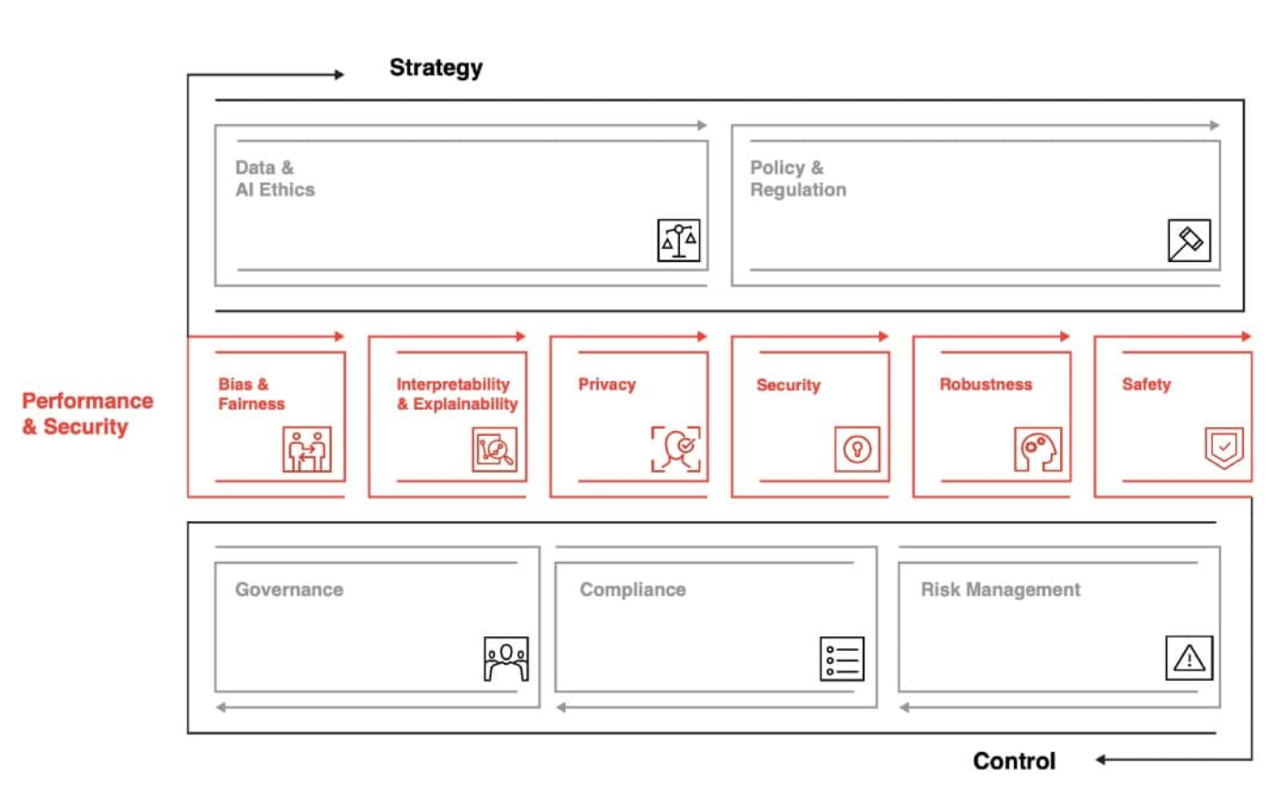
Responsible AI Toolkit: PwC [link]

<https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html>

AI is bringing limitless potential with great risks, and responsible AI (RAI) is the only way to mitigate AI risks. The six major potential risks of applied AI are:

* Performance-errors, instability, bias, opaqueness, lack of interpretability
* Security-adversarial attacks, cyber intrusion, privacy risks, open source software
* Control-lack of human agency, detecting rogue AI, unintended consequences, lack of clear accountability
* Economic-job displacement, enhancing inequality, power concentration within one or a few companies
* Societal-misinformation and manipulation, an intelligence divide, surveillance and warfare
* Enterprise-reputation, financial performance, legal and compliance risks, discrimination, values misalignment

PwC’s Responsible AI Toolkit is a suite of customizable frameworks, tools and processes designed to help to harness the power of AI in an ethical and responsible manner - from strategy through execution, which focuses on and addresses those potential AI risks for organisations.



PwC support all phases of the RAI journey:

* Assess: Technical and qualitative assessments of models and processes to identify gaps
* Build: Development and design of new models and processes, given a specific need and opportunity
* Validate + Scale: Technical model validation and deployment services; governance and ethics change management
* Evaluate + Monitor: Readiness for AI including confirming controls framework design, internal audit training

You will just need to take 10 minutes to complete the PwC’s RAI diagnostic survey. It will generate a score to rank your organisation with actions to consider, which can help you evaluate your organisation’s performance relative to your industry peers, drilling down into questions like:

* How do you protect against attacks and confirm the safety, security and robustness of your AI?
* Are your models treating individuals fairly? Do they respect consumer rights to data privacy?
* Are AI solutions designed and behave in accordance with the law and all relevant regulatory regimes?

**Google – Responsible AI Practices**

Human-centered approach:

* Users appreciate clarity and control – clearly state anything they need to know/you need to disclose
* Consider whether a query is better served by a single response or multiple that the user can choose between
* Attempt to identify user problems/dislikes early in the process. Follow up with testing before full deployment.
* Get input from a broad spectrum users and use cases throughout the development process

Metrics:

* Use multiple forms of feedback (user feedback, system performance, false positive/negative) taken across different contexts
* Use the right metrics for the purpose

Review raw data used for ML training (as far as possible, respecting data sensitivity):

* Mistakes?
* Does sampling represent the end users and usage settings?
* Look for potential deviations between training and application
* Trim unnecessary features
* Is there any difference between ML labels and real variables? Under certain circumstances?
* What biases exist in the data?

Limitations of dataset & model:

* Models can find correlation – doesn’t mean causation
* Coverage of training data limits application range of model
* Explain model limitations to users

Testing:

* Use unit & integration tests on ML components
* Test for input drift
* Test with a high quality dataset to ensure system behaves as normal (and maintain the dataset to cover changing users/use cases)
* Run tests with users for feedback
* Build quality checks into the code

Monitoring & updates:

* Expect issues & plan for problem-fixing time
* Use both short- and long-term fixes as appropriate for the situation
* After updating a model, look for impacts on users and system before deploying

**Microsoft – Responsible AI Practices**

*Principles*

Fairness:

* Important when allocating resources (eg AI for hiring)
* Biases come from all contributors and are difficult to avoid – fairness policing can’t be delegated, diverse team is necessary

Inclusiveness:

* Involve minorities in development and testing
* If it works for the 3%, it probably works for the 97%

Reliability & Safety:

* Systems should minimize errors
* Especially important for physical applications since these can have direct life and death consequences

Transparency:

* A tool to achieve other values
* Understand the reasons behind using AI
* Understand what the AI is doing

Privacy & Security:

* Protect privacy of data sources
* Look for data changes (corruption or adversary attack?)

Accountability:

* Placing requirements on customers/users
* Active assessment of performance against principles

*Accountability Methods*

Impact assessment:

* Dedicated impact assessment required during concept stage
* Review and approve before beginning development
* Update regularly: new userbase, new major release, annually
* Additional classifications when significant adverse impacts are possible: restricted or sensitive use

Fit for purpose:

* Assess during concept stage
* Consider how the system will solve the problem
* For every AI model, explicitly note the inputs, outputs, and limitations
* Determine appropriate performance metrics
* If system is advertised as fit-for-purpose but found to not meet requirements – remove associated uses from customer material and alert current customers

Data governance:

* Document data requirements and collection/processing methods, considering uses and various stakeholders
* Develop data evaluation methods and record evaluation results

Human oversight:

* Identify human overseers responsible for managing AI, and the system elements they need to do so
* Develop metrics for performance of oversight (and use them to evaluate prior to release)

*Transparency Methods*

System intelligibility:

* Identify stakeholders of system outputs (those that use it to make decisions, those affected by the outcomes)
* Design system so these stakeholders able to see the system’s intended uses, its behaviour, and possible bias
* Develop & apply metrics for stakeholder understanding

Communication to Stakeholders:

* Identify stakeholders that interface with the system (deciding to use it or building other integrated systems)
* Develop & publish documentation so these stakeholders understand your system’s capabilities, intended uses, sensitive uses, limitations
* Update documentation regularly and when new users/changes to system/new information

Disclosure of AI Interaction (for systems that fake human interaction):

* Design system so that users know what kind of AI they are interacting with (and label any outputs as such)
* Create metrics to test whether users are aware of the AI interaction (and evaluate performance with them)

*Fairness Methods*

Quality of Service:

* Identify demographic groups at risk of poor service
* Check data for inclusiveness of these groups
* Develop evaluation methods for service to this group (metrics, system tests, data quality) and use these to evaluate system performance
* Where poor service to identified groups is unavoidable, determine reasons
* Publish customer-facing information about groups facing poor performance, the difference in performance, and the identified reasons

Allocation of resources (for systems that influence eg finance, education, employment, healthcare, housing):

* Identify demographic groups at risk of unfair allocation
* Develop evaluation methods for impacts on this group (metrics, system tests, data quality) and use these to evaluate system performance
* Where differences in allocation still exist, determine reasons
* Publish customer-facing information about groups facing impacts, the difference in performance, and the identified reasons

Minimising stereotyping, demeaning, erasing outputs (for systems outputting descriptions or other depictions of people or cultures):

* Identify demographic groups at risk of stereotyping
* Develop a method to assess system components and entire system for stereotyping risks and assess using this method
* Publish customer-facing information about risks and associated demographic groups

*Reliability & Safety Methods*

Reliability & safety guidance:

* Document what acceptable behaviour looks like and acceptable error rates
* Ensure training & test data is representative of intended uses, operation, settings
* Identify ranges for operational variables for general use or specific use cases
* Identify any failure cases in otherwise working ranges
* Document the above & provide details to customers

Failures & remediations:

* Identify predictable failures & their impacts on stakeholders
* For every adverse impact, develop a failure management strategy (redesign system>fallback option) and stakeholder training
* Develop a plan for unforeseen failures

Ongoing monitoring, feedback, evaluation:

* Collate all assessment & monitoring methods for the system, including system analytics, user feedback, and public feedback
* Develop a monitoring plan for each feedback stream – identify important feedback types, event prioritization, customer service input
* Add new feedback streams as required when system/use/etc change – publish info about the support (or lack thereof) for new uses
* Include specific evaluation of Sensitive Use cases (that require unusual qualification or care in use)
* Update documentation when system is found to be not fit for purpose
* Update documentation for new uses, changes to functionality, new information about reliable performance/system accuracy