**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