

# Progressing Towards Responsible AI

Teresa Scantamburlo<sup>1</sup> and Atia Cortés<sup>2</sup> and Marie Schacht<sup>3</sup>

**Abstract.** The field of Artificial Intelligence (AI) and, in particular, the Machine Learning area, counts on a wide range of performance metrics and benchmark data sets to assess the problem-solving effectiveness of its solutions. However, the appearance of research centres, projects or institutions addressing AI solutions from a multidisciplinary and multi-stakeholder perspective suggests a new approach to assessment comprising ethical guidelines, reports or tools and frameworks to help both academia and business to move towards a responsible conceptualisation of AI. They all highlight the relevance of three key aspects: (i) enhancing cooperation among the different stakeholders involved in the design, deployment and use of AI; (ii) promoting multidisciplinary dialogue, including different domains of expertise in this process; and (iii) fostering public engagement to maximise a trusted relation with new technologies and practitioners. In this paper, we introduce the Observatory on Society and Artificial Intelligence (OSAI), an initiative grew out of the project AI4EU aimed at stimulating reflection on a broad spectrum of issues of AI (ethical, legal, social, economic and cultural). In particular, we describe our work in progress around OSAI and suggest how this and similar initiatives can promote a wider appraisal of progress in AI. This will give us the opportunity to present our vision and our *modus operandi* to enhance the implementation of these three fundamental dimensions.

## 1 Introduction

In science, nothing has been more controversial than the notion of progress. Debates on what progress is and if it does really exist abound in the philosophy of science and are closely related to questions about the goals and the methods of a scientific discipline [54]. Although between the 1960s and 1970s philosophers of science put forward thought-provoking views, the idea of progress is commonly associated with the incremental acquisition of knowledge in a particular domain. The same idea is still prevalent in today's research practice, including Artificial Intelligence (AI). For instance, a study discussing the flaws in Google Flu Trends claimed that "*science is a cumulative endeavour, and to stand on the shoulders of giants requires that scientists be able to continually assess work on which they are building*" [47, p 1205].

In this paper, we commit to the view that the progress of a scientific discipline can be measured by the problem-solving<sup>4</sup> effectiveness of its theories [46]. This view applies to, not only science, but

also to different intellectual endeavours, including areas where solutions consist of technical artefacts such as algorithms and computing systems.

Typically, the problem-solving effectiveness of AI solutions is a matter of performance testing. For example, in supervised learning, we assess algorithms based on the number of errors that they make on new, unseen data. Performance lies at the very heart of any learning algorithm, which is, by definition, a computer program that improves through experience.

Intuitively, a Machine Learning (ML) technique is progressive as it performs well along with distinct criteria, including the task at hand, the benchmark data set and the computed performance measure. Although there is no hint of absolute progress in the field - in that no algorithm has proved to be the best at any possible task or condition -, deep learning methods have nevertheless hit the mark in multiple domains ranging from diagnosing eye diseases [37] to playing game [58]. Other signs of progress regard the time of processing [55] which relies on the evolution of CPU capabilities.

However, the introduction of AI algorithms into large portions of human life has suggested that technical performance is not enough. The adequacy of AI solutions depends on a broader set of considerations accounting for the behaviour of AI systems within the environment they are embedded. Repeated facts of algorithmic discrimination and lack of transparency shifted the focus from performance to accountability, from advances in accuracy and speed of computation to the protection of human rights and democratic values. In other words, the appraisal of AI progress is moving away from a purely technical assessment and becoming a multi-factorial affair which integrates aspects of privacy, fairness and transparency, among others.

The transition towards a broader notion of AI assessment is the main focus of the present work. As we will see, a new *ethical turn* prompted the rise of centres and networks addressing AI solutions from a multidisciplinary and multi-stakeholder perspective.

Our main claim is that these initiatives have contributed to moving towards a different notion of progress in AI that goes beyond technical performance to foster responsible development and dissemination of AI, acting on three main elements: (i) promoting an interdisciplinary dialogue around AI; (ii) involving diverse stakeholders and (iii) engaging the public. In particular, we would like to present the Observatory on Society and AI as an example of these multidisciplinary and multi-stakeholder initiatives.

The paper is structured as follows. In Section 2, we describe traditional practices for the assessment of AI, focusing in particular on ML, and the attempts to change them. Section 3 introduces the Observatory for Society and AI as an example of an initiative which may contribute to the *ethical turn* of AI. In Section 4, we survey some responsible practices which will be part of the inventory of resources to be explored by the Observatory. We will conclude in Section 5 with some final remarks.

<sup>1</sup> European Centre for Living Technologies, Ca' Foscari University of Venice, Italy, email: teresa.scantamburlo@unive.it

<sup>2</sup> Barcelona Supercomputing Center, Spain, email: atia.cortes@bsc.es

<sup>3</sup> QuantumBlack, UK, email: Marie.Schacht@quantumblack.com

<sup>4</sup> The notion of problem-solving also describes a specific AI research program (e.g., the General Problem Solver [53]). However, in this paper, this idea refers to the problem-oriented activity of science as neatly described by [46].

## 2 The assessment of progress in AI

One of the keys to understanding the progress of a scientific discipline is to assess its ability to solve problems. Disciplines have developed several methodologies to assess their own problem-solving effectiveness, but what it means for a problem to be solved can vary across the fields. For example, in logic, a problem can be viewed as solved when a theorem has been proved, in philosophy when a thesis is well-argued, in medicine when a treatment cures a disease, and so forth. Each discipline may develop different standards and processes for accepting solutions. The efficacy of a philosophical argument is thus assessed differently from the efficacy of an experimental result.

The field of ML has developed various practices to assess the effectiveness of learning techniques. Most of them reflect the ideal of experimentation in natural science, a model recommended particularly in the early days of ML research [45], then taken as a sign of maturity and objectivity of the field [35]. Usually, testing ML algorithms involve the use of benchmark data sets, the selection of performance measures, multiple tests and comparisons with competing methods. In a classification task, for example, a typical performance metric is accuracy, a scalar value which represents the fraction of predictions that an algorithm got right. More sophisticated methods look at the optimal trade-off between benefits (the true positive rate) and costs (the false-negative rate) such as ROC analysis.

The choice of the performance measure is a crucial task as the available metrics have a distinct meaning which depends on the context of an application. In other words, the measured value represents something we care about [35]. For example, in a system predicting fraud attempts by loan applicants, testing for precision might be sufficient and more informative than other metrics such as specificity. Things would change if the outcome referred to cancer detection, where misclassification comes at different cost.

The creation and maintenance of large data repositories is another influential component of ML testing practice. Since the appearance of UCI collection [28], the field has mostly committed to a benchmark-oriented attitude where data sets from disparate domains become the reference point for comparing algorithms' performance. In recent times, the activity of data collection has witnessed a massive surge and new large-scale databases, as well as more productive data annotation practices, have come to the surface.

The ImageNet project [17] generated more than 14 million images annotated by thousands of crowdworkers and structured around the WordNet hierarchy [29]. ImageNet has also inspired contests (the so-called ImageNet Large Scale Visual Recognition Challenge<sup>5</sup>) where ML practitioners can compete and test their models in different specific tasks, such as object detection and image segmentation. Benchmarks and competitions abound also in natural language processing where we count challenges for question-answering, reasoning and sentiment analysis<sup>6</sup>.

In recent times, the assessment of progress in AI was felicitated by initiatives tracking algorithms' performance during competitions, open repositories and code platforms. These include, for example, the AI index initiative [55] and the AI Watch methodology for the monitoring of AI progress [50]. Thanks to them, one can get a glimpse of the significant breakthroughs achieved by the field. For example, in large-scale object classification tasks, the classification error of

best-performing algorithms fell from 0.28 to 0.023 (see the results presented at CVPR Workshop 2017).

While the systematic analysis of technical performance tells us that the field is progressing at a fast pace, there were AI researchers casting doubts on the robustness of standard assessment approach and claiming that, in reality, what we call progress could be only an illusion [41].

### 2.1 The ethical turn

The discontent with standard testing practice in ML research is not a recent phenomenon. As early as 2006 Chris Drummond [35] criticised the standard testing approach raising three key points:

1. *Performance measures*: he observes that there are other factors influencing a performance measure like accuracy and these may include misclassification costs, the stability of the error rate, and the needs of the end users, among others.
2. *Statistical tests*: following criticism in Psychology, he highlights that statistical tests are frequently misinterpreted, for example, as a confirmation of the alternative to the null hypothesis<sup>7</sup>. While their employment in ML experimental practice is often taken as a sign of rigor and objectivity, he contends that they do not give the degree of evidence that many people believe.
3. *Benchmark data sets*: while the use of benchmark data sets allow us to easily compare algorithms' performance, they suffer from serious limitations. For example, drawing on previous analysis, he raises concerns as to whether benchmark data sets are really representative of reality and how much the data collection and construction account for differences in class distribution.

Further impulse to renovate the experimental practice in ML came from the movement of reproducible research requiring the publication of code and data to reproduce the results reported in scientific articles[59]. In a similar spirit, important ML conferences and journals encourage authors to adhere to best practices by making available data and software tools<sup>8</sup>. Other relevant initiatives include open source repositories and platforms allowing AI practitioners to share data and AI models, such as Papers with Code [26] and Open ML [25]. Another recent work proposed a more comprehensive view of assessment integrating classical performance measures with often neglected costs connected to the development and deployment of an AI system [49].

While these efforts operated within the edges of traditional scientific principles such as transparency and reproducibility, other assessment criteria emerged as a consequence of numerous debates around the impact of AI on society along with plans for action, what we call the *ethical turn* of AI. The new wave of optimism prompted by the successful application of AI to big data [40] was followed by critical analyses raising issues for culture [34, 44] and human decision-making [31]. Other failures and overstatements in the application of AI to transport and healthcare stimulated many initiatives around the world. These comprise research projects, journalistic investigation, centres and networks focused on the impact of AI on our lives.

<sup>7</sup> In statistics the notion of the null hypothesis refers to a default assumption such as "data has a normal distribution" or "there is no correlation between two random variables". In short, a statistical test tells us if we should reject or fail to reject the null hypothesis. In ML the null hypothesis assumes that there is no difference between the performance measures of two algorithms. The one with the greater value is the winner algorithm.

<sup>8</sup> See e.g., the eligibility criteria in the call for paper of the European Conference on Machine Learning (ECML). The Journal for Artificial Intelligence Research has even stricter requirements.

<sup>5</sup> See: <http://www.image-net.org/challenges/LSVRC/>

<sup>6</sup> See e.g., the General Language Understanding Evaluation (GLUE) benchmark (<https://gluebenchmark.com/>), the Stanford Question Answering Dataset (SQuAD) challenge <https://rajpurkar.github.io/SQuAD-explorer/> and the competition section within the CodaLab platform <https://codalab.org/>

In the few last years, we have witnessed more than one hundred declarations of AI principles from governments, organizations and multi-stakeholder initiatives, aimed at providing normative guidance for ethical, rights-respecting, and socially beneficial development and use of AI technologies. Most of those guidelines are aligned on the following eight key themes: Privacy, Accountability, Safety and Security, Transparency and Explainability, Fairness and Non-discrimination, Human Control of Technology, Professional Responsibility, and Promotion of Human Values [38].

Note that, while ethical considerations are not new in the field of AI and notable scholars, such as Norbert Wiener, had already warned of the possible misuse of intelligent and control systems [60], the scale and the number of present efforts have no precedents in the history of the field. For this reason, we acknowledge all these initiatives as a whole movement with the potential to widen future assessment practices.

A final, interesting remark regards the variety of actors who are contributing to this ethical wave. The European Commission, for instance, based on the Trustworthy AI Guidelines (TAIG) [42] published by the High Level expert Group on AI (HLEG-AI), proposed a regulatory framework for high-risk AI applications with a view to build an “ecosystem of trust” [33]. Big companies, likewise, published new design principles<sup>9</sup> and audit frameworks [56], also motivated by practical needs which are usually attenuated in a research context (think of company’s liability and reputation). Finally, the landscape of ethical activities comprises a large number of centres (see Table 1) which scrutinize AI systems through the lens of legal principles and social values, study the impact on economy and human labour, and engage lay people with educational material or works of art.

### 3 An Observatory for Society and AI (OSAI)

The Observatory on Society and Artificial Intelligence (OSAI or simply “Observatory” hereafter) was set up in 2019 within the H2020 EU funded project AI4EU [5], whose objective is to build the first European AI on-demand web platform and ecosystem. The OSAI is an example of this vast array of initiatives animating the *ethical turn* of AI outlined in the previous section. Though at its infancy, it gives us the opportunity to explore how this and similar activities can contribute to stretch the assessment of AI and turns progress towards ethical principles.

The OSAI’s aim is to support discussion and to facilitate the distribution of information about the Ethical, Legal, Socio-Economic and Cultural issues of AI (ELSEC-AI) within Europe. Specifically, the OSAI has the following objectives:

- To stimulate reflection, discussion and due consideration of ELSEC-AI issues within the project through a series of working groups (see Section 4). OSAI is attracting a network of experts in different domains of ELSEC-AI that will contribute to bridge the knowledge gap existing today within AI practitioners and users.
- To provide resources to educate the general EU public more accurately about AI and ELSEC-AI issues by generating weekly content in the form of articles, reports, cultural announcements with the objective to promote discussion and awareness on these topics.

<sup>9</sup> See *e.g.*, the AI Guidelines of Telefónica <https://www.telefonica.com/en/web/responsible-business/our-commitments/ai-principles> or Deutsche Telekom <https://www.telekom.com/en/company/digital-responsibility/details/artificial-intelligence-ai-guideline-524366>

The Observatory evolves in a complex scenario: the field of AI is gaining momentum, and many public and private agencies have begun to consider the opportunities and the risks that lie behind this exciting trend. The OSAI seeks to carve out its own identity and role neither in contrast nor competition with other existing European initiatives (*e.g.*, HLEG-AI). It aims to increase connections among these related projects and make accessible a broad range of articles to the European public at large. The OSAI’s approach can be described by three verbs: 1) *Observe* facts and events occurring within Europe by monitoring newspapers, online bulletins, scientific literature, etc.; 2) *Reflect* on particular events or issues through the contribution of ELSEC-AI experts and, in particular, thanks to the activities of the working groups; 3) *Report* to the general public by using a simple (but not simplistic) language in a way to support mutual understanding among experts and educate lay people.

### 3.1 The context

As we said, the creation of the Observatory takes place in a complex and dynamic context where an imprecise number of AI-related events populate the European calendar. Table 1 includes a collection of European centres that are specifically dedicated to the research around AI and its impact on Society. These were selected from a larger set based on a search of simple keywords on Google engine (such as “AI”, “ethics”, and “society”). The common aim among these institutions is to promote designs and developments of technologies that put upfront concepts such as social responsibility, trust or fairness. Some are dedicated to the creation of guides, others to define evaluation methods, but all have in common the will to create spaces for multidisciplinary dialogue.

While the abundance of centres and projects dealing with AI and its social and ethical impact is a sign of cultural awareness and a source of knowledge, all these positive undertakings run the risk of isolation and self-referentiality. Therefore, OSAI should try to bridge this gap and promote cooperation and mutual knowledge. In addition, it will focus on areas that extend beyond the ethical and legal aspects, including also socio-economic and cultural elements (*e.g.*, how AI is perceived among European citizens, how the arts are presenting or using AI).

The Observatory differs from these initiatives in several respects. In the first place, the OSAI focuses not only on articles and news, but also on people. Indeed, one of the motivating ideas behind the Observatory is the creation of a community of people who can contribute to the discussion of ELSEC-AI. Such a community can combine various types of subjects such as AI experts (*e.g.*, AI researchers and practitioners), specialists in any ELSEC-related field (ethicists, sociologists, lawyers, policy makers, artists, etc) and lay people. In the second place, the OSAI will approach ELSEC-AI in the context of Europe so as to foster the dialogue among European countries.

## 4 OSAI Working Groups

To revise how AI systems are evaluated and complement performance metrics with ELSEC-AI considerations, we need a change in the background of the process. A first step is to promote diversity in the teams that are both designing and assessing such systems. Note that when we talk about diversity, we do not only refer to gender, ethnicity or functional capabilities, but also to include professionals from multiple disciplines and domains of expertise.

A second step is to foster a multidisciplinary dialogue among experts to promote reflection on ELSEC-AI and identify shared strate-

**Table 1.** European centres for AI and Ethics

Name	Country	Type	Objective
HumanE AI [16]	Europe	H2020 EU Project	To create the foundations for AI systems that empower people and society, with special focus on <b>Collaborative Humane Computer Interaction</b> based on a convergence of HCI with ML.
AI Watch [4]	European Commission	Public Institution	An initiative to monitor the development, uptake and impact of AI for Europe
AI4People [6]	European Commission	Multi-stakeholder Forum	To <b>bring together all actors interested in shaping the social impact of new applications of AI</b> , including the European Parliament, civil society organisations, industry and the media. They published the AI4People’s Ethical Framework [39] which inspired the TAIG [42].
OECD.AI [24]	Inter-governmental	International Organisation	The OECD AI Policy Observatory combines resources from across the OECD, its partners and all stakeholder groups to <b>facilitate dialogue between stakeholders</b> while providing <b>multidisciplinary, evidence-based policy analysis</b> in the areas where AI has the most impact.
Knowledge Centre Data & Society [22]	Belgium	Research Centre	Funded by the Flemish Department on Economy, Science and Innovation, it enables <b>socially responsible, ethical and legally appropriate implementations of AI in Flanders</b> .
DataEthics [10]	Denmark	ThinkDoTank	To ensure primacy of the human being in a world of data, based on a European legal and value-based framework. It has a core <b>focus on AI as the evolution of complex data processing</b> extended in human decision-making within politics, economics, identity and culture.
DATALAB - Center for Digital Social Research [12]	Denmark	Research Centre	Conducts research in many different aspects of behavioural data within several areas. A special focus is brought to the <b>social effects of automated data processing</b> as well as to the social adaptation of automated data systems.
ImpactAI [18]	France	Non-profit Association	Think&Do Tank for Ethics and Responsible AI aiming to <b>promote the development of trusted AI</b> , support innovative projects and publish annual reports.
Algorithm Watch [7]	Germany	Non-profit Organisation	Based on research and advocacy to <b>evaluate algorithmic decision-making processes</b> , raise ethical conflicts and explain its features to general audience.
AI & Society Lab [1]	Germany	Research Laboratory	Interface and translator between academia on one side and industry and civil society on the other, it functions as <b>experimental space for new formats to advance knowledge generation and knowledge transfer to AI</b> .
Institute for Ethics in AI [21]	Germany	Research Centre	To generate of global, egalitarian and interdisciplinary <b>guidelines</b> for the ethical development and implementation of AI and to integrate of <b>ethical and societal priorities into the development of fundamentally integrative AI technologies</b> .
AI Sustainability Centre [2]	Sweden	Consultancy	Creation of <b>AI Sustainability Framework</b> for identifying, measuring and governing the ethical implications of AI and assist organisations from a legal, technical and societal perspective.
AI Transparency Institute [3]	Switzerland	Non-profit association	Dedicated to <b>AI governance and human trust in AI</b> , they address key challenges in digital ethics, AI safety, transparency, fairness and privacy.
Digital Ethics Lab [13]	UK	Research Centre	To tackle the <b>ethical challenges of digital innovation</b> from a multidisciplinary perspective, with the aim to identify benefits and positive opportunities while avoiding risks and shortcomings.
Institute for Ethical AI & ML [19]	UK	Research Centre	Highly-technical, practical and cross-functional research across <b>8 Machine Learning Principles and Explainable AI Framework</b>
Institute for Ethical AI in Education [20]	UK	Research Centre	As a response to the TAIG, IEAIE works to <b>develop frameworks and mechanisms</b> to help ensure that the use of AI across education is designed and deployed ethically.
Leverhulme Centre for the Future of Intelligence [23]	UK	Research Centre	To build an interdisciplinary community of researchers with strong links to technologists and the policy world to study the <b>impact of AI in society with a focus on trust, fairness, accountability and democracy</b> .
Centre for Data Ethics and Innovation [8]	UK	Public Institution	Part of Department for Digital, Culture, Media & Sport, they connect policymakers, industry, civil society, and the public to <b>develop the right governance regime for data-driven technologies</b> .

gies to consider ELSEC issues within the AI life cycle. OSAI is trying to fulfill these tasks (diversity and multidisciplinary discussion) by creating a set of working groups, *i.e.*, semi-organised groups of experts working on ELSEC-AI topics.

The experience of the working groups is a laboratory to explore ways to address ELSEC-AI from a diversity of perspectives. They have an experimental character in that there is no systematic knowledge and expertise in dealing with ELSEC-AI in real-world with a multidisciplinary approach. At the beginning, working groups will emphasize two perspectives with a view to incorporating further aspects in future:

- Legal AI: to study existing laws and regulations, how applicable they are to AI systems and identify possible gaps.
- Ethical AI: to promote the design and development of AI systems that respect fundamental rights.

The working groups will be formed by experts in different areas, from academic, business, media or other backgrounds: lawyers and data protection officers, philosophers, software engineers, journalists, sociologists, etc. With this variety of profiles we expect to generate a wide range of opinions and experiences around our topics of interest. Moreover, these working groups are supposed to grow with a bottom-up approach engaging participants from the very beginning. Participants will work in smaller groups and in a limited span time<sup>10</sup>. This activity aims to encourage the participation of experts with an open and transparent methodology and engage them in the communication and sharing of knowledge in this new area of interest. The main objectives of this activity are the following:

- To fill the gap between the ethical debate and the engineering practice in European organizations.
- To help researchers and practitioners navigate the ethical challenges that arise in different real-world AI applications.
- To support interdisciplinary dialogue engaging people from different backgrounds.
- To promote cross-fertilisation among different sectors (*e.g.*, academia, companies, public institutions).
- To inspire future responsible practices in the field of AI.
- To create ELSEC-AI literacy understandable for different types of audiences and domains of expertise.

To fulfill the last two points, we expect to generate a set of good AI practices with a focus on how to implement guidelines such as the HLEG-AI Trustworthy AI Guidelines, especially for SMEs and start-ups. We will also adapt part of the content into educational material that will be shared through the Observatory and the AI4EU communication channels.

## 4.1 Background and Methodology

In the second half of 2019, the authors participated in the piloting of the Trustworthy AI assessment List (TAIL) that was published along with the Trustworthy AI Guidelines by the HLEG-AI, where 50 companies from different sectors and European countries were interviewed. The result of that investigation suggested the need to promote operational resources (*e.g.*, tools, frameworks, procedures/methodologies, etc.) to support organisations to take the Trustworthy AI requirements into account. Other studies have stressed the value to share knowledge and experiences (*e.g.*, best practices, case

<sup>10</sup> Now, the schedule of activities will be constrained by the AI4EU lifetime, but this might be revised in future.

studies) to help practitioners navigate complex ethical issues and interventions [43].

To build upon the piloting of the TAIL, we propose to form small teams that can collaborate with companies (but also public organisations) that are implementing AI products with a view to experimenting how to tackle Trustworthy AI requirements, also drawing on responsible practices (see section 4.2), generate and share a list of good practices that can inspire other organisations. We hope in this way to help fill the gap of knowledge and language between the different stakeholders, *i.e.*, from regulations, to recommendations to the final translation to software engineering methods and other sorts of processes (Human-Computer Interaction methods, ML approaches, management strategies, etc.). Also we will try to disseminate the result of working groups through the Observatory web-site so as to include the Society in this process of gaining trust in new technologies.

At present, the working groups are not intended to generate a comprehensive methodology or a new standard but identify good practices that help integrate ethical and legal requirements into the assessment of AI systems. In concrete, the teams of experts will interact with organisations to analyse specific case studies where they can apply and test one or more responsible AI practices. Interactions can take the form of interviews, design thinking sessions, algorithmic audits, among others, depending on the operational resources adopted by the team (see the responsible practices in Section 4.2). This would help organisations to check whether their AI products meet Trustworthy AI requirements, and possibly re-frame their objectives and their Key Performance Indicator (KPIs) in the light of ethical and legal constraints.

A series of training sessions with invited speakers (either external or internal to the working groups) will help the members of working groups to achieve a shared knowledge about responsible AI practices (questionnaires and checklists, frameworks, strategy guides and canvases, etc). Also, to test and refine our proposal, we plan to pilot working groups with an internal activity involving a few experts and partners (from research and industry) of the AI4EU consortium.

## 4.2 Responsible Practices

While many AI principles have been published, the translation of these into practices and processes is still at the very beginning. The main reason relies on their abstract nature, which makes them difficult for practitioners to operationalise. Even though awareness of the potential ethical issues is increasing at a fast rate, the AI community's ability to take action to mitigate the associated risks is still at its infancy [52]. Operationalisation so far strongly depends on moral compass of individual since there are no legal binds to the existing ethical guidelines. Thus, a framework for ethical decision making and responsible practices is required.

We are classifying the responsible practices by their nature and intended use into 5 groups:

- *Assessments, questionnaires and checklists* raise questions, encourage reflection and inspire potential action, *e.g.*, the HLEG-AI TAIL with 131 questions to operationalise the seven key requirements declared in the AI guidelines [42], or the Consequence Scanning, an agile practice to consider the potential consequences of a product or service on people, communities and the planet [9].
- *End-to-end frameworks* address each stage of the entire process with appropriate activities and involve multiple audiences, *e.g.*, the End-to-End Framework for Internal Algorithmic Auditing to

help companies and their engineering teams audit AI systems before deploying them [56], or the People + AI Guidebook to help user experience (UX) professionals and product managers follow a human-centered approach to AI [27].

- *Strategy guides and canvases* are thinking frameworks and diagnostic tools, that help break down and work through complex challenges, e.g., the Data Ethics Canvas helps identify and manage data ethics considerations [11], or the Ethical Operating System to help inform the design and development process, provide strategies to mitigate risks and take action [15].
- *Design guides* are sets of recommendations towards responsible good practice in design, e.g., the Guidelines for Human-AI Interaction recommend best practices for how AI systems should behave [30]; AI Ethics Cards are a set of four design principles and ten activities that help guide an ethically responsible, culturally considerate, and humanistic approach to designing with data [14].
- *Software toolkits* provide metrics and algorithms to support the ethical development of AI-powered software, e.g., the AI Fairness 360 Toolkit to help examine, report, and mitigate discrimination and bias in ML models throughout the AI application life cycle [32]; Aquitas bias audit toolkit to audit machine learning models for discrimination and bias [57].



**Figure 1.** High-level overview of key pillars for progressing towards responsible outcomes

The landscape of responsible practices is wide, but insufficient, inefficient and scattered. Many of these practices do not adequately address the challenges of context-dependency, they lack ease of use [36] and completeness, they either address only silo disciplines, or single process steps or particular problems. Furthermore, ticking boxes on fairness checklists, mitigating bias with algorithms and anticipating consequences with ethics cards is by far not enough. As depicted in Figure 1, in order to progress towards responsible outcomes, it needs first and foremost close collaboration and a diverse range of perspective, guidance derived from values, principles and policies, a curated set of responsible practices, throughout the entire process, ensured by governance procedures, such as audit services, certifications and AI labels, or company-internal self-check tools, red teams, and ethics Objectives and Key Results (OKRs) and KPIs [51].

As companies start to envision procedures to operationalise AI principles, new professional roles, summarised as “ethics owners” by [51], are being created in order to cover the lack of attention that ethics had in first place. They own responsibility for ethics practices across an organization, and engage to transform principles, val-

ues, ethical stances, and often legal and regulatory imperatives into concrete practices within their organization. To be most effective at achieving these goals, new responsible practices must be aligned with teams’ existing workflows and supported by organizational culture [48].

## 5 Concluding remarks

This paper presents an overview on the existing initiatives that we can find today to enhance the progress of AI towards responsible design, development and deployment. This review puts special attention on the European centres and networks involved in this process since this is the main scope of the Observatory and the objectives within AI4EU. Although it is not exhaustive, it suggests that we are witnessing an *ethical turn* promoting a wider discussion about the assessment of AI systems. In addition, the study on the different responsible practices shows an interest from a diversity of stakeholders to promote a different approach to the assessment of AI, although it is clear that there are still some gaps to cover in order to integrate the required ecosystem.

In his book *Progress and its Problems*, Larry Laudan recommends to cast the nets of appraisal sufficiently widely so as to include *all* the cognitively relevant factors which are *actually* present in the historical situation [46]. The OSAI aims to contribute to this change of paradigm by creating a space for multidisciplinary gathering, where different actors will be able to discuss and create literacy related to ELSEC-AI. We expect to study in depth the key pillars for progressing towards responsible outcomes of AI and their relation with the Trustworthy AI Guidelines and Assessment List in order to find existing relations, but also possible opportunities to complement them. Our final objective is to generate a set of good responsible practices that could help AI practitioners to implement the HLEG-AI documents and align with the European Commission’s vision of Trustworthy AI. We strongly believe that this step is necessary to bring close all the stakeholders involved in the design, deployment and use of new technologies such as AI, which may have great benefits for the Society but are still in a process of being trusted.

## ACKNOWLEDGEMENTS

Teresa Scantamburlo and Atia Cortés are partially supported by the project A European AI On Demand Platform and Ecosystem (AI4EU) H2020-ICT-26 #825619. The views expressed in this paper are not necessarily those of the consortium AI4EU.

## REFERENCES

- [1] AI & Society Lab. <https://www.hiig.de/en/research/ai-and-society-lab/>.
- [2] AI Sustainability Centre. <https://www.aisustainability.org/>.
- [3] AI Transparency Insitute. <https://aitransparencyinstitute.com/>.
- [4] AI Watch. [https://ec.europa.eu/knowledge4policy/ai-watch\\_en](https://ec.europa.eu/knowledge4policy/ai-watch_en).
- [5] AI4EU - A European AI On Demand Platform and Ecosystem. <https://www.ai4eu.eu/>.
- [6] AI4People. <https://www.eismd.eu/ai4people/>.
- [7] AlgorithmWatch. <https://algorithmwatch.org/>.
- [8] Centre for Data Ethics and Innovation. <https://www.gov.uk/government/organisations/centre-for-data-ethics-and-innovation>.
- [9] Consequence Scanning - an agile practice for responsible innovators. <https://www.doteveryone.org.uk/project/consequence-scanning/>.

- [10] Data Ethics. <https://dataethics.eu/>.
- [11] Data Ethics Canvas - Open Data Institute. <https://theodi.org/article/data-ethics-canvas/>.
- [12] DATALAB - Center for Digital Social Research. <https://datalab.au.dk/>.
- [13] Digital Ethics Lab. <https://digitaleticslab.oii.ox.ac.uk/>.
- [14] Ethical Compass - This tool can help. <https://www.ideo.com/blog/ai-needs-an-ethical-compass-this-tool-can-help>.
- [15] Ethical OS Toolkit. <https://ethicalos.org>.
- [16] HumanE AI. <https://www.humane-ai.eu/>.
- [17] ImageNet. <http://www.image-net.org/index>.
- [18] ImpactAI. <http://www.impact-ai.fr/>.
- [19] The Institute for Ethical AI and Machine Learning. <https://ethical.institute/>.
- [20] The Institute for Ethical AI in Education. <https://www.buckingham.ac.uk/research-the-institute-for-ethical-ai-in-education/>.
- [21] Institute for Ethics in AI. <https://ieai.mcts.tum.de/>.
- [22] Knowledge Centre Data & Society. <https://data-en-maatschappij.ai/en/>.
- [23] Leverhulme Centre for the Future of Intelligence. <http://lcfi.ac.uk/>.
- [24] OECD AI Policy Observatory. <https://oecd.ai/>.
- [25] OpenML. <https://www.openml.org/>.
- [26] Papers with Code. <https://paperswithcode.com/>.
- [27] People + AI Guidebook. <https://pair.withgoogle.com/guidebook/>.
- [28] UCI MachineLearning Repository. <https://archive.ics.uci.edu/ml/index.php>.
- [29] WorldNet - a lexical database for english. <https://wordnet.princeton.edu/>.
- [30] S. Amershi, D. Weld, M. Vorvoreanu, A. Fournay, B. Nushi, P. Collison, J. Suh, S. Iqbal, P. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, and E. Horvitz, 'Guidelines for Human-AI Interaction', in *CHI 2019*. ACM, (May 2019). Best Paper Honorable Mention.
- [31] S. Barocas and A. D. Selbst, 'Big Data's Disparate Impact', *California Law Review*, **104**, 671, (2016).
- [32] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, 'Ai Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias', *IBM Journal of Research and Development*, **63**(4/5), 4:1-4:15, (2019).
- [33] European Commission. White Paper on Artificial Intelligence: a European approach to excellence and trust. [https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust\\_en](https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en), 2020.
- [34] d. boyd and K. Crawford, 'Critical Questions for Big Data', *Information, Communication & Society*, **15**(5), 662-679, (2012).
- [35] C. Drummond, 'Machine Learning as an Experimental Science (revisited)', in *Proceedings of the Twenty-First National Conference on Artificial Intelligence: Workshop on Evaluation Methods for Machine Learning*, pp. 1-5. AAAI Press, (2006).
- [36] AI Ethics Impact Group, *From Principles to Practice - An interdisciplinary framework to operationalise AI ethics*, Bertelsmann Stiftung, 2020.
- [37] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, B. O'Donoghue, D. Visentin, G. van den Driessche, B. Lakshminarayanan, C. Meyer, F. Mackinder, S. Bouton, K. Ayoub, R. Chopra, D. King, A. Karthikesalingam, C. O. Hughes, R. Raine, J. Hughes, D. A. Sim, C. Egan, A. Tufail, H. Montgomery, D. Hassabis, G. Rees, T. Back, P. T. Khaw, M. Suleyman, J. Cornebise, P. A. Keane, and O. Ronneberger, 'Clinically Applicable Deep Learning for Diagnosis and Referral in Retinal Disease', *Nature Medicine*, **24**(9), 1342-1350, (2018).
- [38] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI'. Berkman Klein Center for Internet Society, (2020).
- [39] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, 'Ai4people - An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations', *Minds Mach.*, **28**(4), 689-707, (December 2018).
- [40] A. Halevy, P. Norvig, and F. Pereira, 'The Unreasonable Effectiveness of Data', *IEEE Intelligent Systems*, **24**(2), 8-12, (2009).
- [41] D. J. Hand, 'Classifier technology and the illusion of progress', *Statistical Science*, **21**(1), 30-34, (2006).
- [42] High-Level Expert Group, *Ethic Guidelines for Trustworthy AI*, European Union, 2019.
- [43] K. Holstein, J. Wortman Vaughan, H. Daumé, M. Dudik, and H. Wallach, 'Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?', in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 1-16, New York, NY, USA, (2019). Association for Computing Machinery.
- [44] R. Kitchin, 'Big Data, new epistemologies and paradigm shifts', *Big Data & Society*, **1**(1), 2053951714528481, (2014).
- [45] P. Langley, 'Machine Learning as an Experimental Science', *Mach. Learn.*, **3**(1), 5-8, (August 1988).
- [46] L. Laudan, *Progress and Its Problems: Towards a Theory of Scientific Growth*, University of California Press, Berkeley, CA, 1978.
- [47] D. Lazer, R. Kennedy, G. King, and A. Vespignani, 'The Parable of Google Flu: Traps in Big Data Analysis', *Science*, **343**(6176), 1203-1205, (2014).
- [48] M. A. Madaio, L. Stark, J. Wortman Vaughan, and H. Wallach, 'Co-designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI', in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, p. 1-14, New York, NY, USA, (2020). Association for Computing Machinery.
- [49] F. Martínez-Plumed, S. Avin, M. Brundage, A. Dafoe, S. hÉigeartaigh, and J. Hernandez-Orallo. Accounting for the Neglected Dimensions of AI Progress. <https://arxiv.org/abs/1806.00610>, 2018.
- [50] F. Martínez-Plumed, E. Gomez, and J. Hernandez-Orallo, 'AI Watch Methodology to Monitor the Evolution of AI Technologies'. Publications Office of the EU, (2020).
- [51] J. Metcalf, E. Moss, and D. Boyd, 'Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics', volume 82, pp. 449-476, (2019).
- [52] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal. From what to how: An initial review of publicly available ai ethics tools, methods and research to translate principles into practices, 2019.
- [53] A. Newell, J. C. Shaw, and H. A. Simon, 'Report on a General Problem-Solving Program', in *Proceedings of the International Conference on Information Processing*, pp. 256-264, (1959).
- [54] I. Niiniluoto, 'Scientific Progress', in *The Stanford Encyclopedia of Philosophy*, ed., Edward N. Zalta, Metaphysics Research Lab, Stanford University, winter 2019 edn., (2019).
- [55] E. Brynjolfsson J. Clark J. Etchemendy B. Grosz T. Lyons J. Manyika S. Mishra R. Perrault, Y. Shoham and J. C. Niebles, 'The AI Index 2019 Annual Report', Technical report, Human-Centered AI Institute, Stanford University, (December 2019).
- [56] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, 'Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing', in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT\* '20, p. 33-44, New York, NY, USA, (2020). Association for Computing Machinery.
- [57] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani. Aequitas: A Bias and Fairness Audit Toolkit, 2018.
- [58] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, 'Mastering the game of go without human knowledge', *Nature*, **550**, 354-359, (2017).
- [59] S. Sonnenburg, M. L. Braun, C. S. Ong, S. Bengio, L. Bottou, G. Holmes, Y. LeCun, KR Müller, F. Pereira, C. E. Rasmussen, G. Rätsch, B. Schölkopf, A. Smola, P. Vincent, J. Weston, and R. Williamson, 'The Need for Open Source Software in Machine Learning', *J. Mach. Learn. Res.*, **8**, 2443-2466, (2007).
- [60] N. Wiener, *The Human Use of Human Beings*, Houghton Mifflin, Boston, MA, 1954.