Responsible AI through a Software Engineering lens @ Serlab

Maria Teresa Baldassarre, Vita Santa Barletta, Danilo Caivano, Domenico Gigante, and

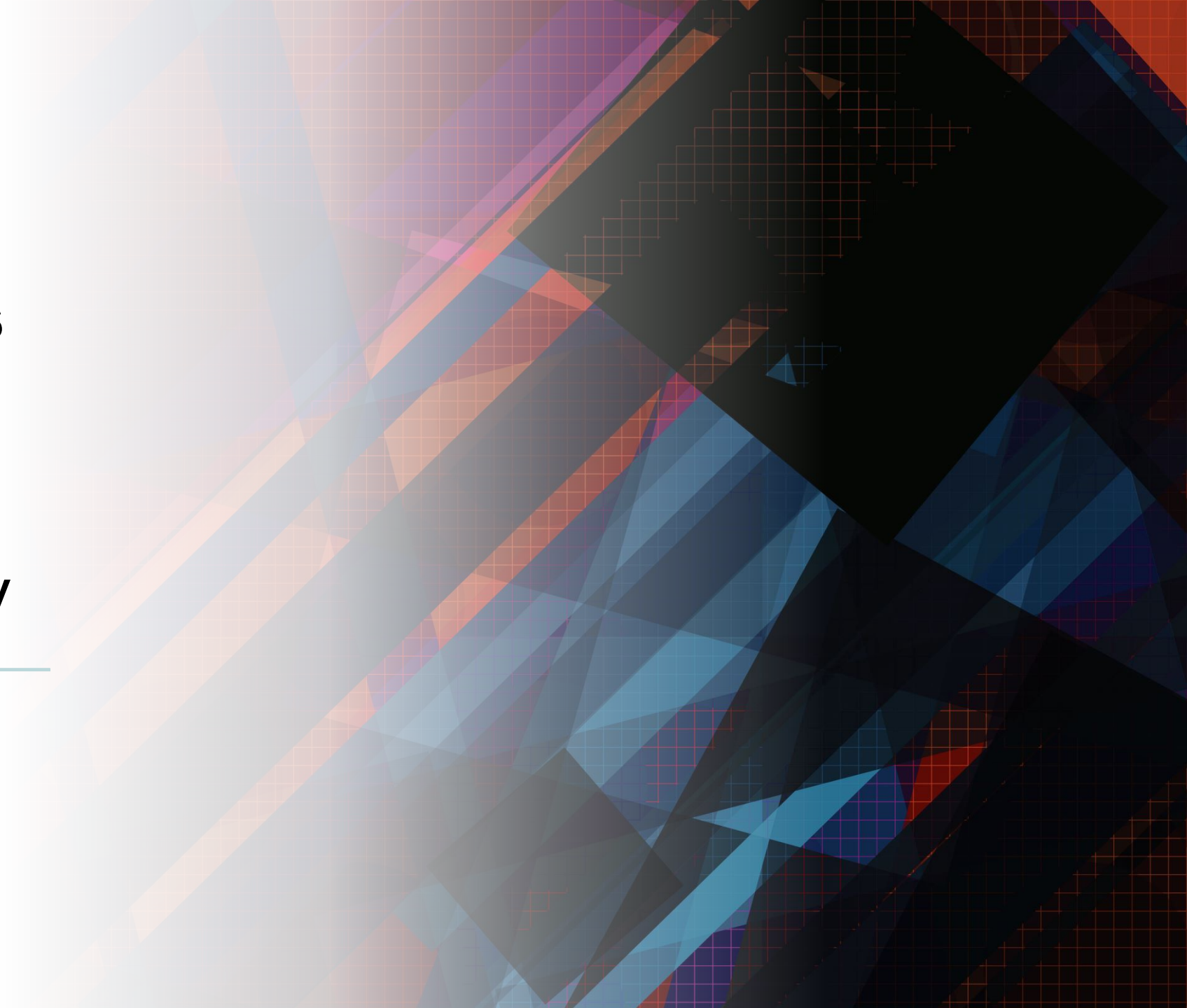
Azzurra Ragone

UNIVERSITY OF BARI

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

How to guide the development of ethical AI

What are the **ethical principles** that should lead the AI **implementation** and **use** in society



Our research goal

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Study what **AI practitioners**, both technical and non-technical stakeholders, **need** in term of guidelines, best practices, and tools to be **supported** and **guided** in the development and deployment of Responsible AI applications in all the Software Development Lifecycle (SDLC).

Research Questions

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

RQ1: What is the **state of the practice** and the correlated literature to approach the Responsible AI development?

RQ2: **What** do the **practitioners think** about Responsible AI? What are their **perceived gaps**?

RQ3: Is it possible to **realize a framework** able to support **different kinds of stakeholders** in implementing Responsible AI?

Roadmap

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Rapid review in the field of Responsible AI, to understand what has been done, gaps and needs

→

Study (survey) directly with the **practitioners**, to understand their real needs and validate the results obtained from the rapid review

→

Development of a **framework prototype** to guide different stakeholders (technical and non-technical) in the development of RAI applications

Responsible AI principles

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

To address the problem of **principle proliferation**, we have decided to focus on a specific subset of principles

The four principles identified by Jobin et al. [13]:

transparency, justice and fairness, non-maleficence, responsibility, and privacy

with the exclusion of **responsibility** as this concept is rarely defined in a clear manner.

Responsible AI principles definition

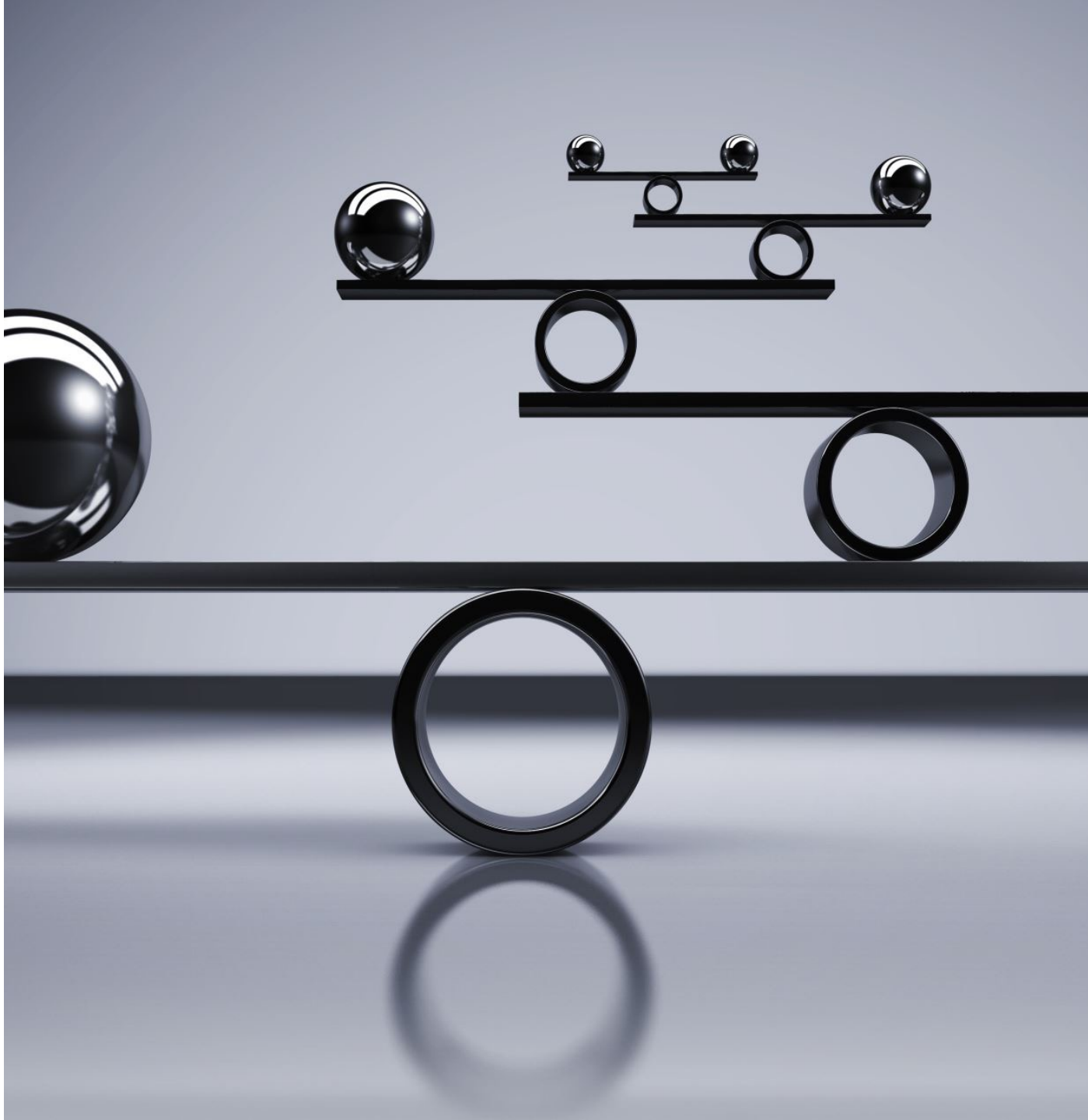
The chosen principles are:

Transparency (known also as *explainability*)

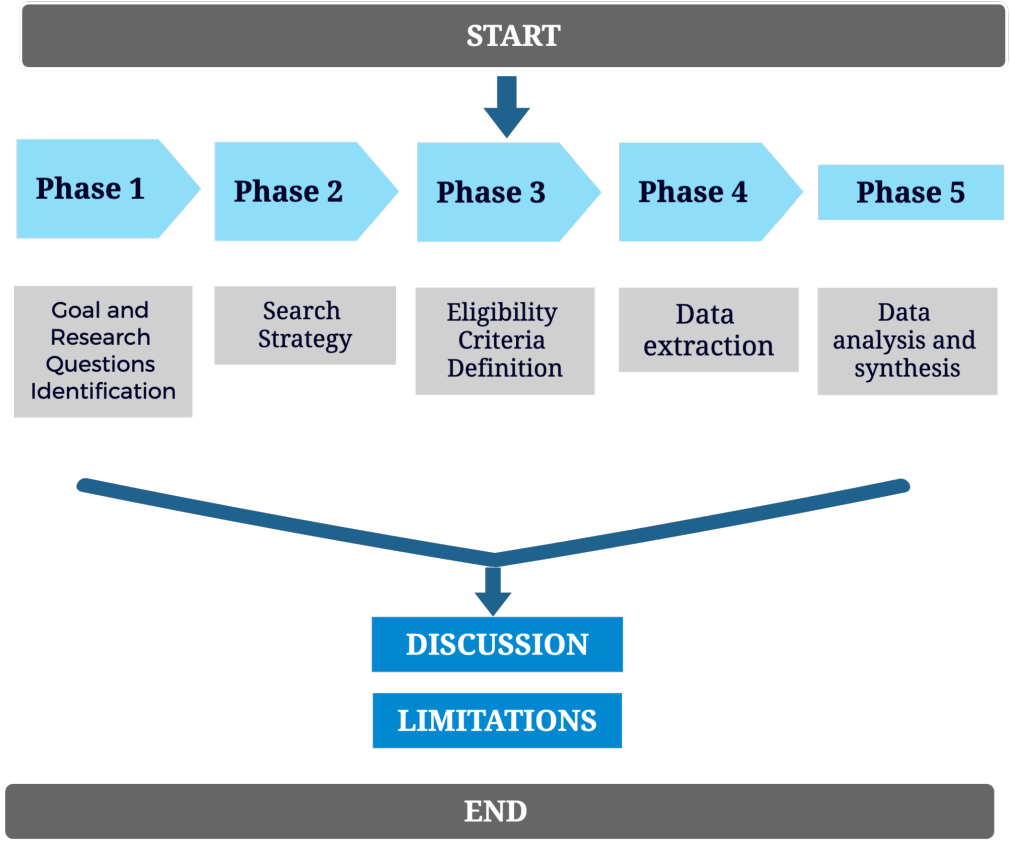
Diversity and non-discrimination and fairness (as *Justice and fairness*)

Technical robustness and safety (as *Non-maleficience*)

Privacy and data governance



A Rapid Review of Responsible AI frameworks



Research Questions

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

RQ1: What are the Responsible AI frameworks proposed in the literature?

RQ2: How much do these frameworks address the various RAI principles?

RQ3: Do these frameworks provide recommendations for each phase of the Software Development Life Cycle (SDLC)?

RQ4: Is there a supporting tool for each proposed framework?

Search strategy

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

White literature search: Scopus, Google Scholar

Grey literature search: Algorithm Watch, OECD database, Google search engine

Eligibility criteria definition

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

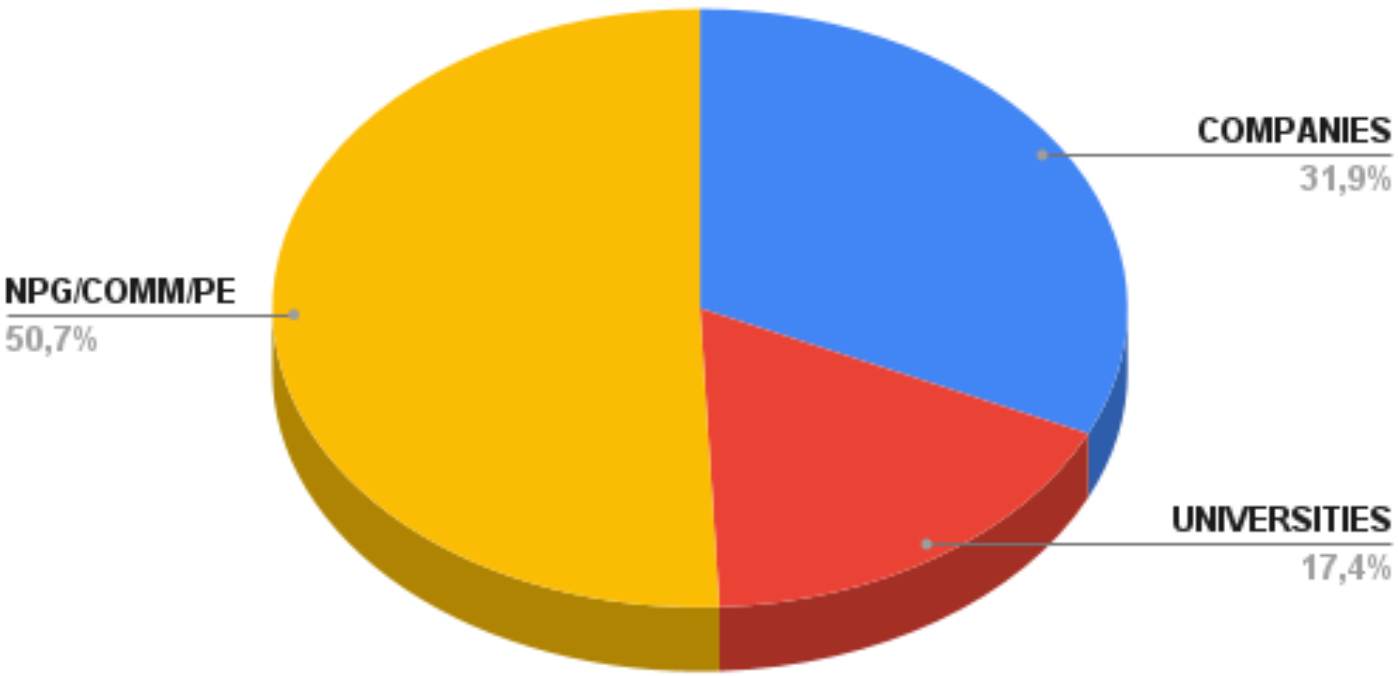
1. The resource must be in English or Italian
2. The resource must be in the context of Responsible AI frameworks
3. The resource must address at least one of the chosen principles
4. The resource must provide answers to at least one of the rapid review's research questions.

Data extraction

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Data Source	Resources retrieved	Resources analyzed	Resources selected
Scopus	1875	1489	20
Google Scholar	91200	200	0
Algorithm Watch	167	167	80
OECD DB	356	70	38
Google Search	2110000	168	10

Results



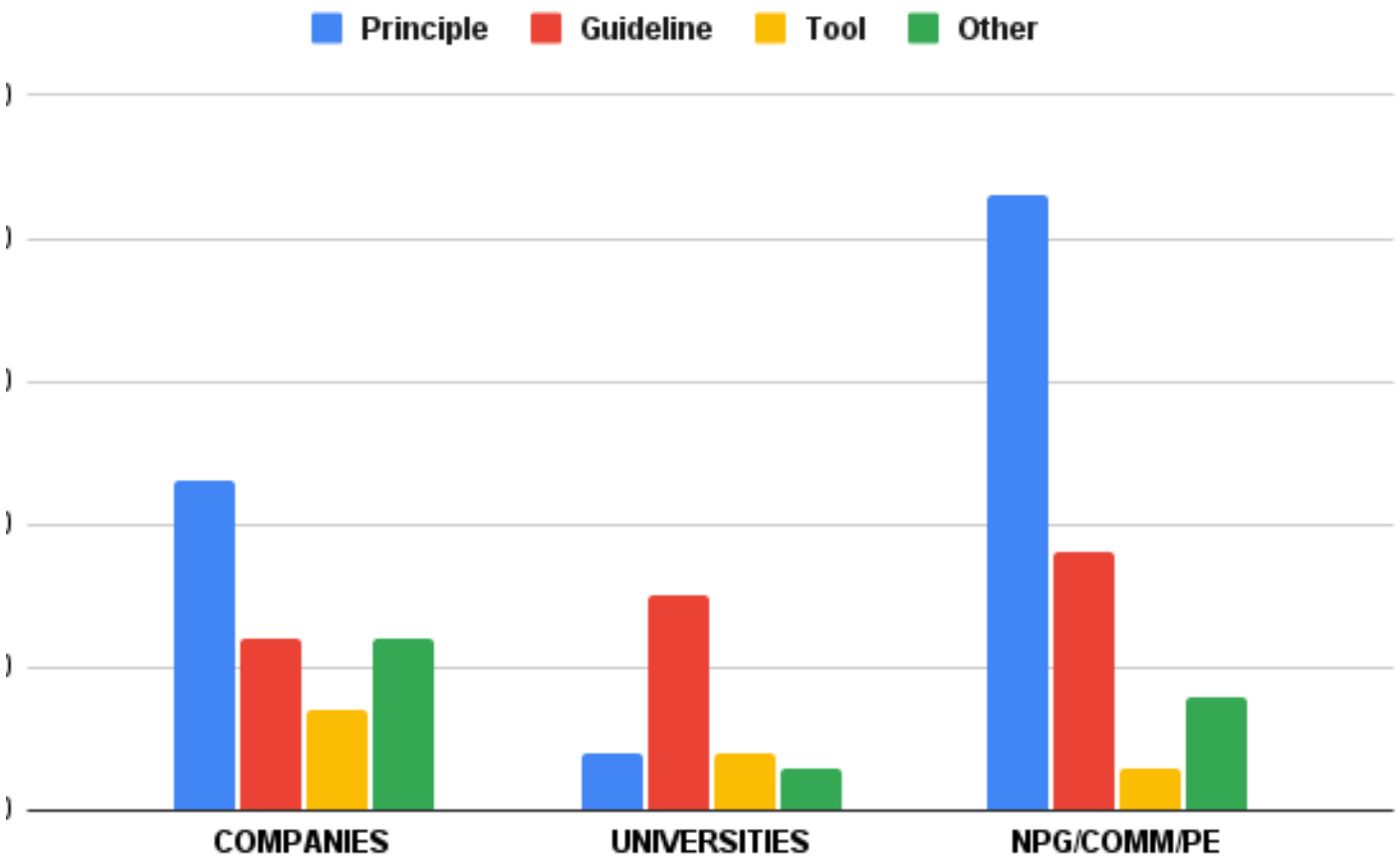
All the retrieved frameworks have been classified w.r.t. the type of proposing institution

Framework classification

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

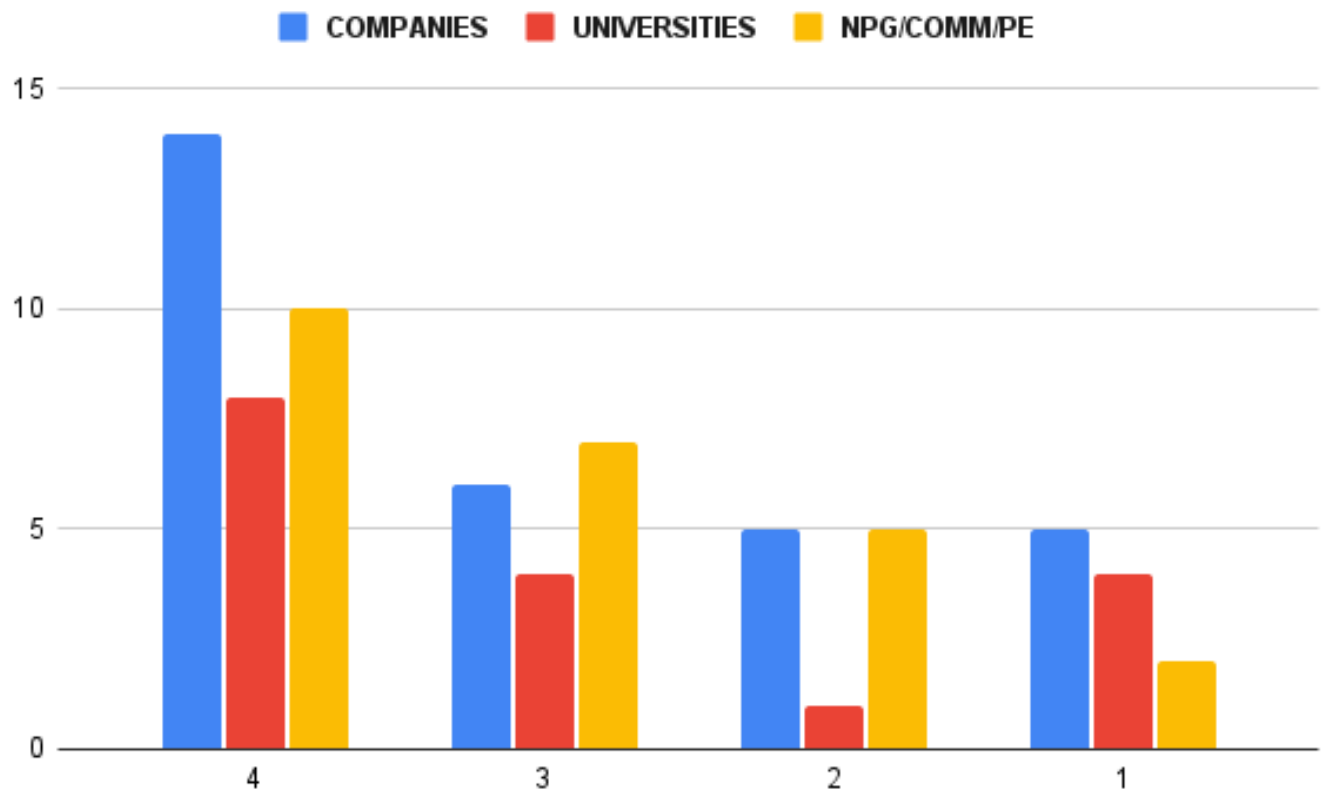
1. **Principle** (P): highlight only abstract ethical principles or moral values;
2. **Guideline** (G): concrete guidelines, quickly translatable into design constraints or choices;
3. **Tool** (T): verify the compliance towards one or more principles and/or support practitioners in the implementation of principles or guidelines;
4. **Other** (O): if a resource cannot be classified into any of these categories

Results



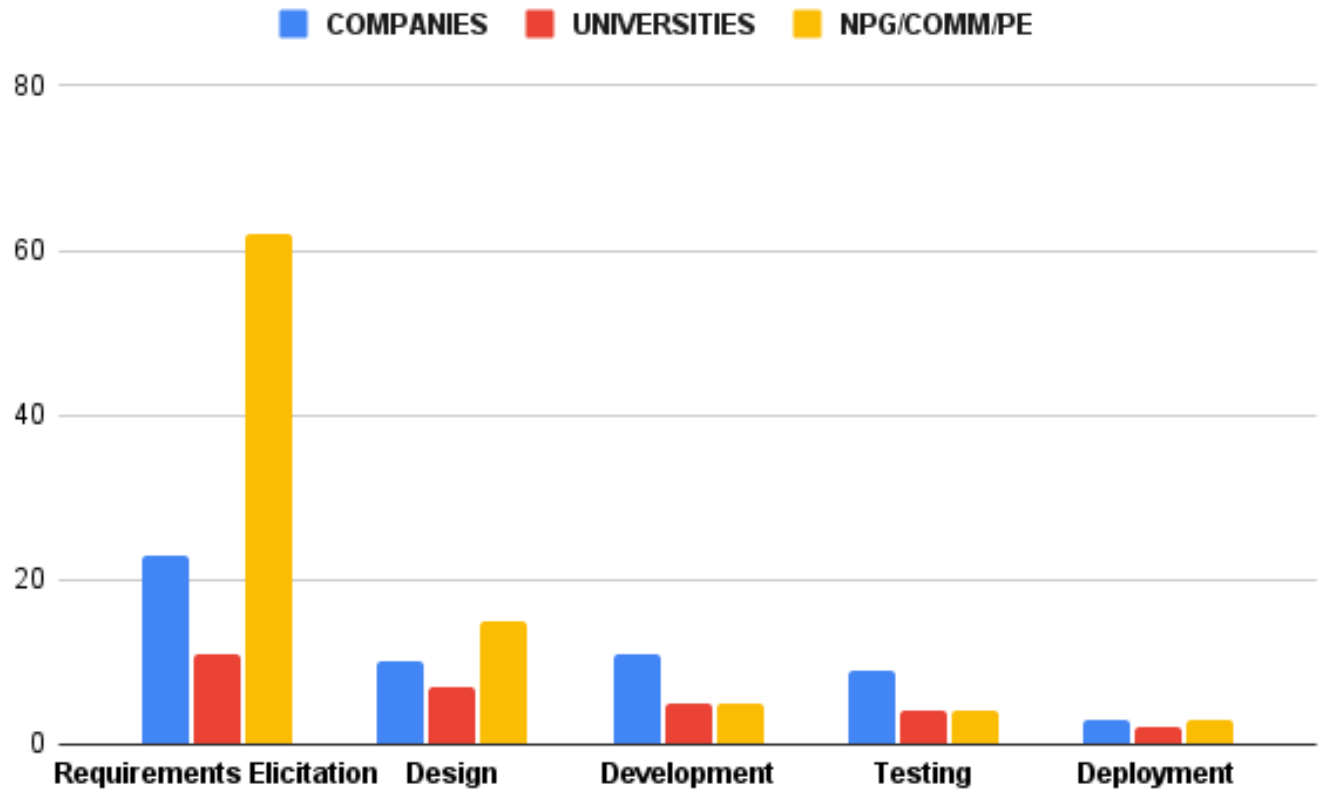
The distribution of frameworks by their category and grouped by proposing institution

Results



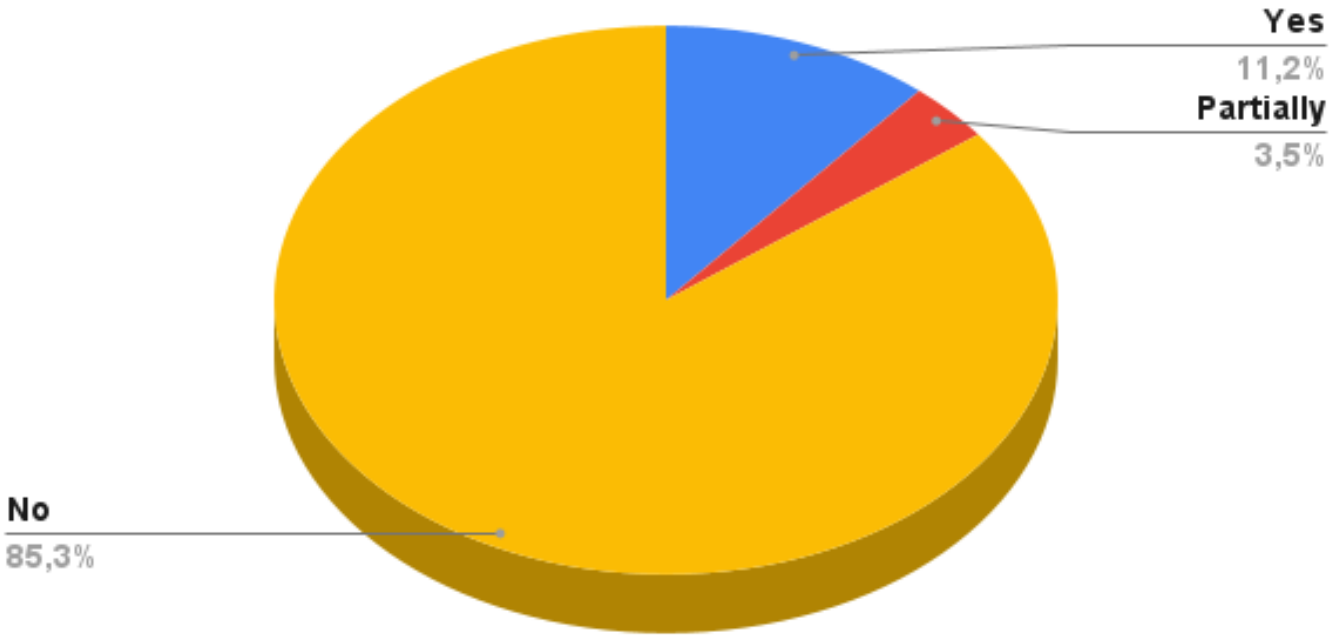
Number of RAI principles addressed by the frameworks grouped by proposing entity type.

Results



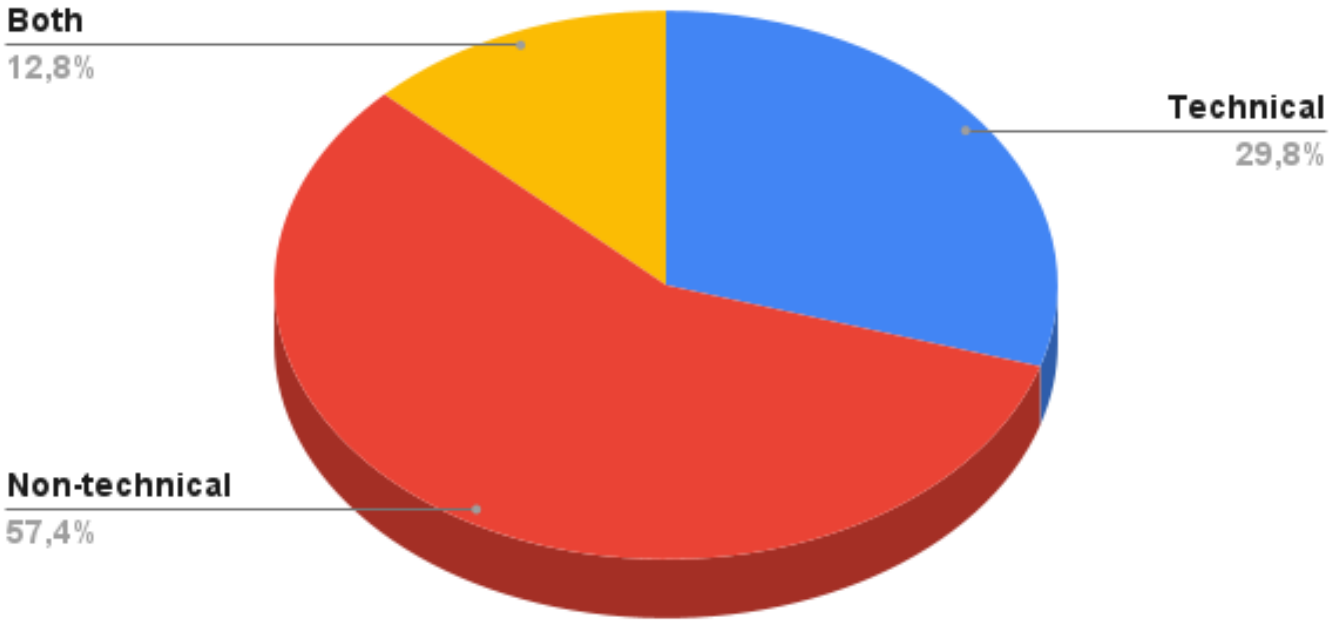
Distribution by SDLC phase addressed.

Results



Distribution by the existence of a supporting tool regardless of the proposing entity

Results



Stakeholder’s required background regardless of proposing entity in case a tool is provided.

Current practice gaps

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

- Most of the filtered frameworks are **proposed by No-profit organizations / Communities / Public entities** (50,7%). Regarding the type, we can say that there is a **worrying lack of tools**: most of the frameworks are just Principles or Guidelines.

- The **majority of the frameworks address all four principles** previously presented, sometimes in a "*partial*" way: this reveals an even greater **lack of consensus and standardization** about which are the best practices to follow to be compliant with the RAI values.

- **Very few frameworks encompass all the SDLC phases**; most frameworks focus only on the **initial phases of the SDLC**, and, specifically, on *Requirements elicitation*.

- In most **cases there is not a practical tool complementing the theoretical frameworks**; this is true regardless of the type of entity releasing the tool.

Rapid Review Wrap up

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

To summarize, **right now does not exist any comprehensive framework whose knowledge can be navigated and exploited by different kinds of stakeholders** (technical and non-technical ones), which can simplify and speed up the adoption of RAI practices.

Ongoing work

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Our next step consists in **spreading a survey among AI experts** (both from industry and academia) to collect as much structured data as possible, to derive an initial preview of the **actual practical gaps in the state of the practice** and to **extract the key points requiring a deeper investigation**.

Then we want to analyze these key points **by conducting focus groups** in which we ask the practitioners if they agree regarding the gaps that emerged from literature on Responsible AI.

This formalized data will enable us to answer **RQ2**.

Bibliography

A rapid review of Responsible AI frameworks: How to guide the development of ethical AI. The International Conference on Evaluation and Assessment in Software Engineering (EASE) - 2023

Responsible AI through a Software Engineering lens @ Serlab

Maria Teresa Baldassarre, Vita Santa Barletta, Danilo Caivano, Domenico Gigante, and

Azzurra Ragone

azzurra.ragone@uniba.it