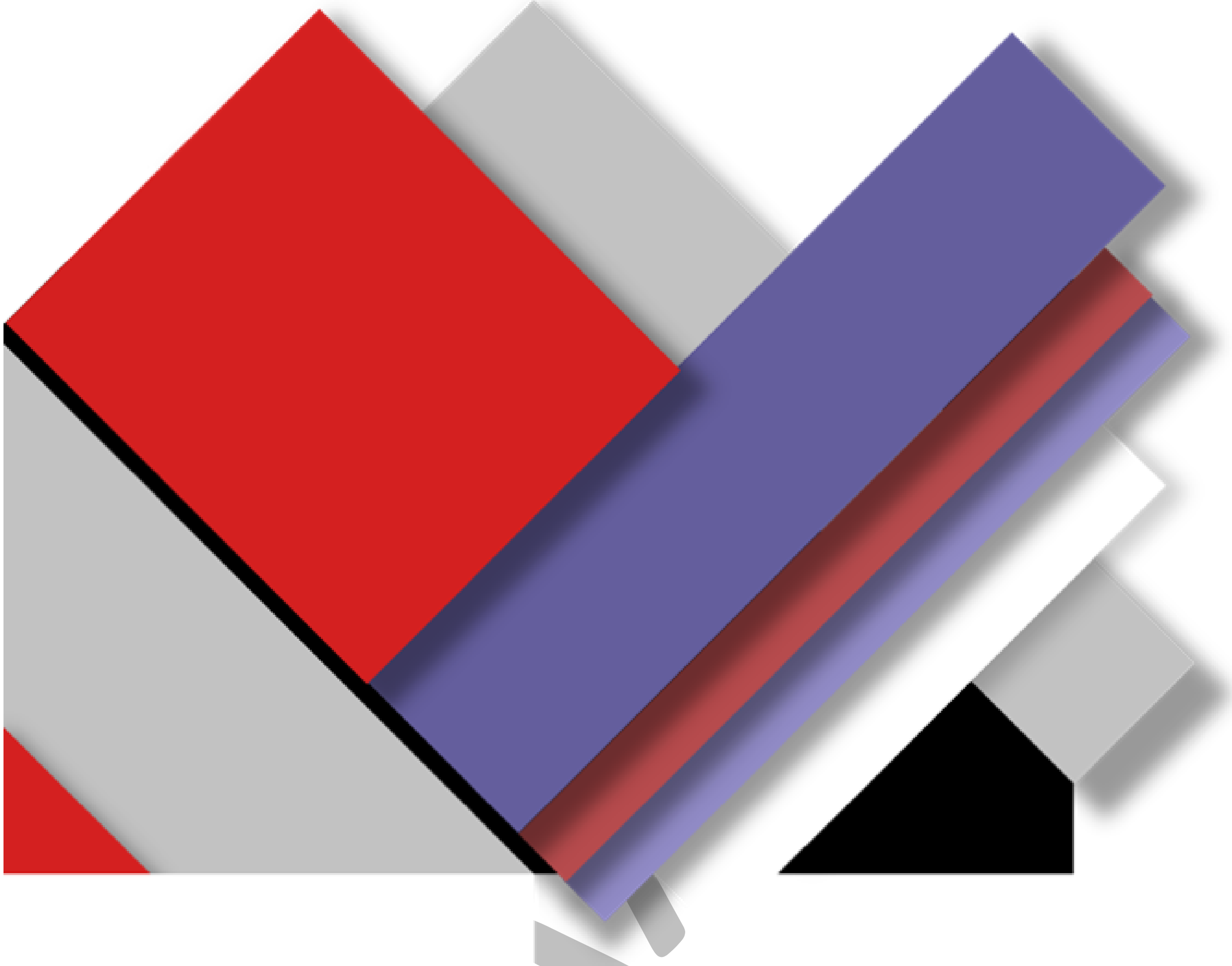

Kairoi – Technical Background

25th July, 2022

Kairoi – Ismael Kherroubi Garcia

London, UK

Overview

This document outlines the technical research that underpins Kairoi's services. It explains how the relevant research is *interpreted* and later *implemented*.

Naming Kairoi

Kairoi (pronounced */kye-roy/*) is from the Ancient Greek word for *opportune time*. It is not the same as the term *kronos*, which describes time as following an orderly sequence. Rather, *kairos* describes moments when decisions are critical. Indeed, time-allocation is a question of ethics. As we learn from *Harvard Business Review*:

> "People tend not to think of allocating time as an ethical choice, but they should. Time is a scarce resource, and squandering it—your own or others'—only compromises value creation. Conversely, using it wisely to increase collective value or utility is the very definition of ethical action" (Bazerman, 2020).

At Kairoi, we believe most decisions in tech have the potential to lead to great social impacts. We help our partners identify these crucial decisions, anticipate their consequences and implement safeguards to guide decision-making processes.

This Report's Structure

There are dozens – even hundreds – of frameworks on the ethics of artificial intelligence (AI) technologies (Dotan, 2021). Whilst the number of relevant documents will continue to grow and inform our practices at Kairoi, we have identified four works that shape our approach to AI ethics and research governance:

1. Douglas (2014) *The Moral Terrain of Science*
2. Burget et al. (2017) *Definitions and Conceptual Dimensions of Responsible Research and Innovation*
3. Jobin et al. (2019) *The Global Landscape of AI Ethics Guidelines*
4. Wong et al. (2022) *Seeing Like a Toolkit: How Toolkits Envision the Work of AI Ethics*

These research papers are interpreted in the following section *§Technical Background*. The first three resources are then mapped onto a practical canvas for guiding responsible approaches to research and innovation. This is available in *§The RRI Canvas*. The canvas is supplemented with guidance on each of its sections.

Whilst the canvas and its guidance are made available for anybody to read, share and apply to their own professional contexts (this document is made available under a CC BY-NC-SA 4.0 licence), the section *§Implementation* suggests the stages through which Kairoi can help you make the most of the canvas.

Technical Background

This section contains summarised interpretations of the four most influential papers on Kairoi's approach to AI ethics and research governance. In short, Douglas (2014) describes the philosophical foundation underpinning *The RRI Canvas*, Burget et al. (2017) introduce concepts that appear in the practice of RRI; Jobin et al. (2019) list organisational practices to implement moral values in RRI practices; and Wong et al. (2022) describe limitations to ethics frameworks that we must avoid.

Douglas' Moral Terrain of Science

Philosopher Dr. Heather Douglas outlines in her 2014 paper a framework for evaluating responsibility in scientific communities. Whilst not all of Kairoi's partners will see themselves as conducting science proper (whatever that means), they will be following rigorous methods and modern techniques to develop novel insights and technologies. To this effect, a framework intended for science can be of use for thinking about responsible research and innovation across industries.

Douglas identifies the following four bases for scientific responsibility:

- Scientific reasoning "requires of scientists a concern for genuine empirical discovery", and responds to both:
    - The scientific community ("it is in a robust intellectual community that scientists can be assured of doing their best work"), and
    - Broader society. In this sense, the interaction of societal and scientific values must:
        - Be made explicit
        - Be transparent, particularly when societal values override scientific endeavours
        - Targeted, so as to not undermine the value of science (i.e.: ban specific practices, rather than broader objectives).

- The nature of the responsibility, which can be:
    - General ("responsibilities we all share by virtue of being human, by the fact that we are part of a human community"), or
    - Role-specific ("responsibilities we take on when we adopt a particular role").
- The level of responsibility:
    - Minimum demand: "Minimum standards are floors, and they need to be firm and clear, with mechanisms of rebuke when one falls below them. Minimum standards should be things like 'don't steal someone else's work, cite work that you quote, don't make up evidence, don't inadvertently kill research assistants,' and so forth."
    - Ideal: "scientists should (and often claim to be) trying to make society better, to help humanity, and to improve the state of the planet."
- Who bears the responsibility:
    - The individual (whose responsibilities can be collectivised when complex)
    - The community (via irreducibly social governance structures)

As Douglas was referring to scientific practice, we must reinterpret the framework to make it more relevant to the diverse organisations working on technological innovations. The first step for doing so is to dampen the emphasis on "the scientific community" brought forth by the first basis. With this, we can allow for more diverse communities to be considered when assessing the rigour or robustness of scientific work. Indeed, we can bring more diverse views into discussions about technological advancement by doing so.

Continuing with the first basis, Douglas' apparent distinction between societal and scientific values is a complex topic far beyond the scope of this report. Nonetheless, it is worth noting that, at Kairoi, we assume our partners strive to follow the most cutting-edge, accurate and appropriate scientific tools and methodologies, but this can be covered by the second basis.

The nature of the responsibility can be either *general* or *role-specific*. In practice, the latter has already been touched upon in a specific sense: in doing their jobs *well*, researchers must operate to the standards of their respective scientific communities. *Role-specificity*, in the case of Kairoi, means identifying the practical responsibilities of as many jobs within an organisation as possible. For example, researchers are responsible for following the standards of their relevant *epistemic communities*, and human resources personnel are responsible to their own agreed upon standards, and so on. In the simplest sense, we can take role-specific responsibilities to be those reflected in job descriptions.

General responsibility is implemented through agreed upon belief systems informed by our positions within the social world. These no longer differ from job to job but are chares regardless of your position within an organisation. When speaking of implementing general responsibilities, we can find a great deal of relevant policies to be inspired by. Corporate Social Responsibility (CSR) statements or Diversity and Inclusion (D&I) policies are just some ways to create an organisational culture whereby each individual agrees to adhere to certain moral standards.

Douglas' two *levels of responsibility* constitute a philosophically complex feature within the framework. The idea is that there are two extremes one can be responsible for: either leave the world no worse than you found it, or do your best to make it better than it was. In practice, we can interpret these extremes as either producing the least demanding outputs (for example, responding to your emails) or going above and beyond (morally speaking, that is; for example, by being mindful of your colleagues' wellbeing).

Note that, the moment we list examples of either extreme, debates can emerge: "is answering *all* your emails ever the point?" "How much effort should I put into my colleagues' wellbeing?" To this effect, at Kairoi, we interpret Douglas' levels of responsibility as describing the extremes of a *spectrum*. With our clients, we help identify both extremes and expect practices to sit somewhere in between. But by mapping *ideal behaviours*, we can identify ways for organisations to improve and always strive for better.

Two types of *agents* can bear responsibility according to Douglas: *individuals* and *collectives*. In organisations, identifying individuals should be quite straightforward: they are our employees, contractors, visitors, and so on. Conversely, collective agents include *formal groups* within an organisation's structures. For example, there may be groups of investors, trustees or other oversight committees to whom senior staff report. There may also be different groups of "diversity champions" who advocate for improvements in staff mental health, gender and racial equity, or disability inclusiveness. Many other collective decision makers may exist. The point is that these groups each hold their own types of responsibility.

Conceptual Dimensions of Responsible Research and Innovation

Burget et al. (2017) identify and define four established conceptual dimensions of RRI from an analysis of relevant policies. These are:

- Inclusion: Uncovering societally desirable outcomes through participatory approaches to science;
- Anticipation: Envisioning the future of research and innovation and understanding how current dynamics can help design the future, and avoid potentially harmful consequences from technologies;

- Responsiveness: Risk identification, and transparent and accessible scientific results; and
- Reflexivity: Public dialogue, science and public collaboration, anticipation, and bringing social scientists and philosophers into lab deliberations.

In the same study, Burget et al. identify *sustainability* and *care* as emerging dimensions in RRI practices. These – in total – six dimensions serve as inspiration for identifying *ideal behaviours* as defined by Douglas above. They should not be seen as values that can be perfectly fulfilled, but practices to strive for. Here are examples of ideal behaviours that the six dimensions might inspire:

- Inclusion: Conduct focus groups and other citizen science practices, with a particular eye to historically marginalised communities;
- Anticipation: Partner in government initiatives to legislate AI technologies;
- Responsiveness: Follow *open science* practices to help broader research communities advance technological knowledge
- Reflexivity: Foster an interdisciplinary organisational culture that values *employee voice*
- Sustainability: Employ renewable energy and conduct environmental impact assessments; and
- Care: Gain insights into the public's belief systems and lived experiences when deploying participatory science practices.

Eleven Values in AI Ethics and Four Reasons for Divergent Interpretations

Jobin et al. (2019) conducted a review of 84 documents containing guidelines for AI. Their analysis found eleven moral values that overlapped in many of those documents. They also found that the values were not interpreted equally across documents. This is not surprising, as it is difficult to imagine there being universally valid conceptions of such abstract ideas. Specifically, they identified four reasons why the values might be interpreted differently. The eleven values are listed below, and the four reasons are briefly introduced:

Values	Reasons they may be interpreted differently
Transparency	**Definition:** The basic notion of the value may differ. For example,

Justice & fairness	*responsibility* might refer to either the need to hold humans accountable for technologies gone awry, or the need to view humans as responsible for technological artefacts from the outset.
Non-maleficence	
Responsibility	**Justification:** The reason why a value is believed to matter may differ. For example, *transparency* might be deemed important because it is a way to foster public trust, or because transparency is seen as a way to mitigate any harms.
Privacy	
Beneficence	
Freedom & autonomy	**Application domain:** Who or what processes a value is relevant to might be seen differently. *Non-maleficence* might refer to the potential for a technology to be of *dual-use* (which might not sit well with some), or the idea that unintended harms are unavoidable (such that mitigation strategies are always necessary but never sufficient).
Trust	
Sustainability	
Dignity	**Implementation:** Each value can be implemented through different mechanisms. For example, for some, *sustainability* can be implemented by improving the energy efficiency of AI tools and research methods; for others, it is a case of ensuring that there is accountability for AI technologies that lead to job losses.
Solidarity	

A table containing the eleven values and how each is reinterpreted for each of the four reasons is available in *§Appendix 1*.

Importantly, at Kairoi, we strive to implement tangible mechanisms for change. Thus, what is most interesting from Jobin et al.'s work are those diverse methods for implementing values. *§Appendix 1* contains a summary table of the eleven values and four reasons for divergent interpretations that Jobin et al. identify. The table also includes a label for each implementation method. This goes beyond what Jobin et al. provide, but it shows how most of the methods overlap. The four overlapping categories of value-implementation mechanisms are the following:

- Communications strategies
- Technical solutions
- Participatory approaches
- Governance

These categories are not exhaustive, and can overlap in many instances, but they can serve organisations to think about the gaps they have to better reach for *ideal behaviours*.

How Toolkits Envision the Work of AI Ethics

In a preprint, Wong et al. (2022) outline the assumptions and limitations they find in 27 AI ethics toolkits. Two types of assumptions are generally embedded in such toolkits; assumptions about who is responsible for AI to be ethical, and about what the work of "doing AI ethics" entails. Roughly, they find the following assumptions about each:

- Who does the work:
    - Engineers and data scientists must envisage and implement specifications
    - Executives must be convinced of the commercial value of responsible tech
    - AI design and development teams are ill-defined
    - Non-technical staff are seen as important but only vaguely, and they are not provided with any resources to adequately engage in these conversations
    - Power imbalances are rarely mentioned, let alone remedied
    - External groups whose roles are also underspecified include: "clients, vendors, customers, users, civil society groups, journalists, advocacy groups, community members, and others impacted by AI systems"
- What the work entails:
    - Technical work
    - Engagement with stakeholders external to the team or company
    - Finding solutions

Wong et al. conclude with the following three broad recommendations:

- Embrace the non-technical dimensions of ethics work
- Support for engaging with stakeholders from non-technical backgrounds
- Structure the work of AI ethics as a problem for collective action

These recommendations are not dissimilar from what we find in the aforementioned analyses. Technical solutions are one category of the mechanisms identified by Jobin et al. (2019), and the need for working with "non-technical" people is captured by Burget et al.'s (2017) *reflexivity*. Meanwhile, the need to see AI ethics as a job for collective action rather than something that needs "solving" can be framed by Douglas' identification of *collective responsibility*.

With these four foundational documents outlined, we can turn to implementing their insights through *The RRI Canvas*.

The Kairoi Canvas

The Kairoi Canvas is theoretically informed, evidence-based, and academically rigorous. It can be seen as containing three parts. The first is informed by Douglas’ (2014) *Moral Terrain of Science*. It straightforwardly draws on three of the four bases for responsibility in science: the nature or type of responsibility, the responsible agent or party, and the extreme levels of responsibility, from *minimum demands* to *ideal behaviours*. *Ideal behaviours*, in turn, can be informed by the six dimensions of RRI: *inclusion*, *anticipation*, *responsiveness*, *reflexivity*, *sustainability* and *care* (Burget et al., 2017).

Responsibility	Minimum Demands	Ideal Behaviours
Role-specific for individuals		
Role-specific for collectives		
General		

The second part of *The RRI Canvas* is informed by Jobin et al.’s (2019) analysis of AI ethics frameworks. Kairoi’s independent reanalysis of their work shows that there are at least four categories of options to implement responsible practices in research and innovation organisations.

From Minimum Demands to Ideal Behaviours			
Communication strategies	Technical solutions	Participatory approaches	Governance

Finally, *The RRI Canvas* highlights the need for any implementation mechanism to be embedded through continuous training. Thus, the template canvas available to all to work with looks like the following:

<table>
<tr><th>Responsibility</th><th colspan="2">Minimum Demands</th><th colspan="2">Ideal Behaviours</th></tr>
<tr><td>Role-specific for individuals</td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>Role-specific for collectives</td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td>General</td><td colspan="2"></td><td colspan="2"></td></tr>
<tr><td colspan="5">From Minimum Demands to Ideal Behaviours</td></tr>
<tr><td>Communication strategies</td><td>Technical solutions</td><td colspan="2">Participatory approaches</td><td>Governance</td></tr>
<tr><td></td><td></td><td colspan="2"></td><td></td></tr>
<tr><td></td><td></td><td colspan="2"></td><td></td></tr>
<tr><td></td><td></td><td colspan="2"></td><td></td></tr>
<tr><td colspan="5"><b>Training</b></td></tr>
</table>

Responsible Technology Strategy

Understanding the theoretical underpinning of *The RRI Canvas* is insufficient itself to make a real impact. Furthermore, the canvas itself is only a tool or *heuristic device* to support broader strategies for organisational and social change. At Kairoi, we want to see our clients thrive by doing what's best for them and for the world. By identifying *minimum demands* and *ideal behaviours*, we can map their processes and identify mechanisms to develop more responsible technologies. This allows us to develop and implement new and improved practices that feed into a much broader *Responsible Technology Strategy*.

The four main stages that Kairoi follows to help you make the most of *The RRI Canvas* and design a Responsible Technology Strategy are summarised below. Of course, as with all organisational change strategies, the method will adapt to the particular context of our clients, including their specific sectors and available resources.

1. The status quo

    a. Identify organisational stakeholders (departments and power dynamics) and staff (organigram).
    b. Outline each party’s responsibilities (“role”).
    c. Outline current practices.

2. A reflexive future
    a. Become comfortable discussing moral values
    b. Identify, define and communicate organisational values.
    c. Apply these values to identify “minimum demands” and “ideal behaviours”.
    d. Assign responsibilities in light of the stakeholder-mapping exercise (i.e.: Who is responsible for each type of behaviour, and is this role-specific or general?)
    e. Map current practices onto the framework (i.e.: Do they meet minimum demands, or sit somewhere between these and ideal behaviours?)

3. Time for improving
    a. Identify gaps between current practices and ideal behaviours.
    b. Establish the most feasible modes of action to fill in the gap (pick from comms, technical solutions, participatory approaches and governance structures).
    c. Implement through training (at induction and through upskilling of current staff).

4. Monitoring
    a. Establish a mechanism that updates the list and outline of stakeholders, current practices and responsibilities.
    b. Update minimum demands and ideal behaviours as organisational values and social context change.
    c. Conduct exercise (2) periodically.

Appendix 1

Categorising value-implementation mechanisms

The following table contains Jobin et al’s (2019) eleven values and four reasons for divergent interpretations. Each implementation mechanism is categorised as *communication strategy*, *technical solution*, *participatory approaches* or *governance*. These categories are not exhaustive, and can overlap in many instances, but they can serve organisations to think about the gaps they have to better reach for *ideal behaviours*.

	Interpretation	**Justification**	**Application domain**	**Implementation**
Transparency	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing	● A way to minimise harm and improve AI ● Legal benefits ● Fosters trust ● Links to dialogue, participation and democratic principles	Data-use, human-AI interaction, automated decisions, the purpose of data use or application of AI systems	● Disclosure of information by AI systems developers and deployers **[Communication strategy]** ● Communicate use of AI, source code, data use, evidence-base for AI use, limitations, laws, responsibility for AI, investments in AI, and/or possible impacts **[Communication strategy]** ● Provision of non-technical explanations **[Communication strategy]** ● Audits ● Technical solutions **[technical solutions]** ● Oversight, interaction and mediation with stakeholders and the public, and the facilitation of whistle-blowing **[participatory approaches]**

Justice, fairness and equity	● Fairness, prevention, mitigating unwanted bias and discrimination ● Justice as respect for equality, diversity and inclusion ● Possibility to appeal or challenge decisions ● Right to redress and remedy ● Fair access to AI, data and their benefits ● AI's impact on labour market ● Need to address democratic or societal issues ● Need to acquire and process accurate, complete and diverse data, especially training data			● Technical solutions (standards or explicit normative encoding) **[technical solutions]** ● Transparency by providing information and raising public awareness of existing rights and regulation **[Communication strategy]** ● Testing, monitoring and auditing ● Developing or strengthening the rule of law and the right to appeal, recourse, redress or remedy **[governance]** ● Systemic changes and processes (governmental action and oversight, a more interdisciplinary or otherwise diverse workforce, better inclusion of civil society or other relevant stakeholders in an interactive manner, and increased attention to the distribution of benefits **[governance]**
Non-maleficence	● Safety and security ● Never cause foreseeable or unintentional harm ● Avoid specific risks (e.g.: misuse for cyberwarfare and malicious hacking) Harm as.. ● Discrimination, violation of privacy, bodily harm		● Potential dual-use ● Unavoidable damages whereby risks should be assessed, reduced, and mitigated ● Technical and governance strategies applied to: ● AI research ● AI systems design	● Risk-management strategies ● Technical solutions include in-built data quality evaluations, security or privacy by design, and establishing industry standards **[technical solutions]** ● Governance strategies include active cooperation across disciplines and stakeholders, compliance with existing or new legislation, and oversight pratices tests, monitoring, audits, assessments by

	● Loss of trust or skills, radical individualism, the risk that technological progress outpaces regulation ● Negative impacts on long-term social wellbeing, infrastructure or psychological, emotional or economic dimensions		● AI systems development and deployment ● Continuous integration	internal unites, customers, users, independent third parties, or governmental entities) **[governance]** ● Clear definitions of liability-attribution policies **[governance]**
Responsibility and accountability	Responsibility, accountability, liability, acting with integrity Hold AI accountable in a human-like manner vs. always deem humans as responsible for technological artefacts		AI developers, designers, institutions, industry	● Act with "integrity" ● Clarify attribution of responsibility and legal liability (up-front, in contracts or by centering remedy) **[governance]** ● Focus on underlying reasons and processes that may lead to harm ● Whistleblowing policies, promotion of diversity, introduction of ethics to STEM education **[governance]**
Privacy	Privacy, personal or private information ● A value to uphold and/or a right to be protected ● Data protection or security ● Freedom or trust			● Technical solutions (differential privacy, privacy by design, data minimisation, access control) **[technical solutions]** ● Calls for more research and awareness ● Regulatory approaches (legal compliance, certification, more specific legislation) **[governance]**
Beneficence	Benefits, beneficence, well-being, peace, social good, common good		● Customers ● Everyone, humanity, society, as many	● Align AI with human values ● Advance scientific understanding of the world

	● Augmentation of human senses ● Promotion of human wellbeing and flourishing ● Peace and happiness ● Creation of socioeconomic opportunities ● Economic prosperity		people as possible, all sentient creatures ● The planet and the environment	● Minimise power concentration ● Use power “for the benefit of human rights” **[governance]** ● Work closely with “affected” people ● Minimise conflicts of interest **[participatory approaches]** ● Prove beneficence through customer demand and feedback **[participatory approaches]** ● Develop new metrics for human wellbeing **[governance]**
Freedom and autonomy	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment ● Specifically freedom of expression, informational self-determination or privacy-protecting user controls ● Freedom, empowerment or autonomy ● Positive freedom to flourish, to self-determination through democratic means, to establish and develop relationships with other human beings, to withdraw consent, or to use a preferred technology			● Transparent and predictable AI ● Do not reduce options for and knowledge of citizens **[communication strategy]** ● Increase people’s knowledge about AI **[communication strategy]** ● Give notice and consent **[communication strategy]** ● Refrain from collecting and spreading data in absence of informed consent

	• Negative freedom from technological experimentation, manipulation or surveillance			
Trust	Trust and trustworthiness • Trust as AI being transparent, understandable or explainable • Trust as AI fulfilling public expectations	A culture of trust among scientists and engineers is believed to support the achievement of other organisational goals, or because overall trust in the recommendations, judgments and uses of AI is indispensable for AI to ‘fulfil its world changing potential’ But pay attention to “excessive trust in AI”	Trustworthy… • AI research and technology • AI developers and organisations • Design principles	• Education **[communication strategy]** • Accountability • Processes to monitor and evaluate the integrity of AI systems over time • Tools and techniques ensuring with norms and standards **[technical solutions]** • Certificate of fairness **[governance]** • Multi-stakeholder dialogue **[participatory approaches]** • Awareness about the value of using personal data and avoiding harm **[communication strategy]**
Dignity	Dignity as intertwined with human rights, which means avoiding harm, forced acceptance, automated classification, unknown human-AI interaction	AI should not diminish or destroy but respect, preserve and increase human dignity	A prerogative of humans but not robots	Dignity is preserved if: **[governance]** • Respected by AI developers in the first place, and promoted through… • New legislation, • Governance initiatives and • Government-issued technical and methodological guidelines

Sustainability	Sustainability, environment (nature), energy, resources (energy) ● Protecting the environment, improving ecosystem and biodiversity ● Contributing to fairer and more equal societies ● Promoting peace		Process data sustainably and provide insights that remain valid over time	● Improve AI systems' efficiency and minimise ecological footprint **[technical solutions]** ● Policies that ensure accountability in the domain of potential job losses, and use challenges as an opportunity for innovation **[governance]**
Solidarity	Solidarity, social security, cohesion		Implications of AI for the labour market Avoid radical individualism	● Strong social safety net **[governance]** ● Redistribute benefits of AI to not threaten social cohesion and respect vulnerable persons or groups **[governance]**

Appendix 2: Example Canvas

Responsible Party	Responsibility Type	Minimum Demands	Ideal Behaviours
Individuals			
Principal Investigators	Role (job-related)	Communicate research goals	Foster collaboration within your team and across teams
	General (human)	Do not leave the world worse than you found it	Inclusiveness, responsiveness, reflexivity
Project Managers	Role (job-related)	Coordinate resources	
	General (human)	Do not leave the world worse than you found it	Make the world a better place
Etc.			
Collectives			
Executive leaders	Role (internal consistency; SWOT)	Financial viability	Fair wages, open-mindedness
	General (PESTLE)	Legal compliance, human rights	Anticipation, sustainability
Etc.			
From Minimum Demands to Ideal Behaviours			
Communication strategies	Technical solutions	Participatory approaches	Governance structures
Methodological limitations	Differential privacy	Stakeholder management	Industry standards
Explanations	Auditing & monitoring	Co-design	Research ethics committee
Legal rights	Improved efficiency	Customer feedback	Impact assessment(s)
Training			