This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

Draft awaiting European Commission approval

GRANT AGREEMENT NUMBER

101035819

This project has received funding from the European Union's
Horizon 2020 research and innovation programme

ENLIGHT RISE- RESEARCH AND INNOVATION AGENDA WITH AND FOR SOCIETY: LEVERAGING DIGITAL INNOVATION FOR A GREENER AND HEALTHIER EUROPE

WP No	Del. Rel. No	Del No	Title	Lead beneficiary
3	D3.4	D20	D3.4 White Paper for an ENLIGHT Responsible DI/AI Index	UU

Nature	Dissemination Level	Related to Del. No (if applicable)
Report	Public	

Description (short)
This report will describe how ENLIGHT can actively contribute to advancing responsible DI/AI R&I, to optimize its impact on the 5 flagship challenges and develop a responsible DI/AI Development Index to assess and monitor the implementation of these guidelines.

Document Version Control Draft awaiting European Commission approval		
Version 0.1	Lead author: Sébastien Peyrard, Charlotte Cosin, Margaretha Andersson	UBx (lead), UU
Version 0.2	Lead author: Sébastien Peyrard, Charlotte Cosin, Margaretha Andersson Contributors: Merle Schatz, Birgit Schmidt, Philip Van Den Heede, Grégoire Sierra	UBx (lead), UU, UGOE, RUG
Version 1.0	Lead author: Sébastien Peyrard, Charlotte Cosin, Margaretha Andersson Contributors: Merle Schatz, Birgit Schmidt, Philip Van Den Heede, Grégoire Sierra All Enlight partners (AI in education survey)	UBx (lead), UU, UGOE, UGhent
DOI	https://doi.org/10.5281/zenodo.10889332	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

Draft awaiting European Commission approval

Table of contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

Draft awaiting European Commission approval

Responsible Innovation: Context, Definitions and Framework

The idea of science serving the greater good is perhaps as old as science itself, with famous examples François Rabelais's "Science without conscience is but ruin of the soul"[1]. This question taked on a particular resonance in an era marked by climate change and the rise of generative AIs and what looks like a new industrial revolution. This relates to the concept of Responsible Research and Innovation, how it can be enabled, fostered and sustained in the Anthropocene. The concept of responsible research and innovation is as "a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view on the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society)" (European Commission. Directorate-General for Research and innovation & Schomberg, 2011). Ideally, a set of principles and indicators in a DI/AI responsible index would enable innovators to assess the responsible nature of their DI/AI innovations, and the wider public to evaluate them.

This notion of responsible research and innovation is an apparent contradiction as "sustainable development relies on having a long-term vision, taking into account general interest, the precautionary principle, the shared risks and benefits between stakeholders while "market" innovation is based on the capacity of an enterprise to propose quick - or first - new solutions in order to increase its share of economic benefits") (Marcandella, 2015).

What's more, can we assess innovation in itself as "good" or "bad"?

Innovation for good? Limits and caveats

Can we judge an innovation *per se* as responsible or not?

"Tech for good" is not a very relevant way of dealing with the problem for several reasons. It is inherently impossible to assess an innovation on these grounds because of several principles (Silberzahn, 2020):

- **Uncertainty**: one can never foresee the uses to which a technology will be put. Many technologies have been initiated in one field and then reused in other disciplines and other sectors.
- **Ambiguity**: technology can be used for good or bad purposes.
- **Subjectivity:** the definition of what is a "good" or a "bad" use depends on the socio-cultural context and cannot be assessed in isolation.
- **Necessity**: it is not necessary to want to do the good in order to actually do it.
- **Risk not-to-innovate**: the above principles show that preventing innovation in the face of a "bad" effect can prevent innovation that has the potential to bring huge, unforeseen benefits to society.

(Silberzahn, 2020) gives the example of ultrasound, invented in 1911 to detect submarines, used in medicine in the 1950's in echography with enormous benefits for health and well-being, then used in the 1980's for sex-selective abortions. More recently, the end of constitutional protection for abortion in the US raised serious concerns about how data from period-tracking smartphone apps

[1] « Science sans conscience n'est que ruine de l'âme ». François Rabelais, *Gargantua*, 1542.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

could be used in court cases in states where abortion is considered a crime (Kelly & Habib, 2023). This example shows how an innovation aimed at empowering women can have quite the opposite effect in a different context. As can be seen from such examples, it is impossible to evaluate an innovation *per se*. What's more, it shows that the evaluation of a technology cannot be done at a single point in time, but should be a continuous process throughout its lifecycle: before conception (does this meet an existing need?), during conception (trying to anticipate possible external effects during its lifecycle), after release (what are the indirect, unanticipated effects that need to be considered?), as defined by Xavier Pavie (Pavie, 2012). As Sauzet puts it, "responsible innovation management aims to integrate innovation throughout its process of emergence, dissemination and reuse, as well as individuals or groups impacted directly or indirectly by innovation" (Sauzet, 2022)

The Collingridge[2] dilemma shows the limits of such a view: "Ethical issues could be easily addressed early on during technology design and development whereas in this initial stage the development of the technology is difficult to predict" (European Commission. Directorate-General for Research and innovation & Schomberg, 2011). In other words, there is "a double-bind problem:

- An **information problem**: impacts cannot be easily predicted until the technology is extensively developed and widely used.
- A **power problem**: control or change is difficult when the technology has become entrenched." ("Collingridge dilemma," 2023).

This dilemma highlights the ambiguity of the term "innovation" in itself, as it can refer to both the outcome and the process. While the former meaning lends us to think of an innovation as a finished object, the latter insists on the fact that innovation is an ongoing process, suggesting continuous monitoring and reviewing rather than a one-off assessment of a product (at what stage? With what use? etc.).

Responsible Innovation Management and Normative Background

In the light of such conundrums, it is more relevant to talk about Responsible Innovation *Management*, than "just" responsible innovation in itself. Societal Responsibility assessment is a socio-technical issue, as it is related to political and organizational matters, not merely technical or even scientific ones. According to (Fernez-Walch & Romon, 2016), "innovation cannot be "responsible" in itself on a societal or environmental ground. But an organization can, as an economic stakeholder, be responsible on a societal level". Marcandella comments this by adding that "one cannot assign responsibility to an object, service or process whereas one can assign responsibility – be it individual or collective – to the stakeholders of an innovation process" (Marcandella, 2015).

This nuance is reflected in ISO 5600X norms such as ISO 56002 (International Organization for Standardization, 2018), that have additional guidelines in the AFNOR standard FD X50-271, which consider innovation management as an organizational process and mindset, rather than a one-size-fits-all recipe to be applied to a project.

[2] Named after David Collingridge who initiated the dilemma in his 1980 book *The Social Control of Technology.*

If we add the "societal responsibility" layer to the problem, we may turn to another related set of norms, the ISO 26000 standard (International Organization for Standardization, 2020) and e.g. its AFNOR FD X30-031 (Association Française de Normalisation, 2013) implementation guidelines.

This set of norms provides several insights into the assessment of social responsibility in innovation projects.

The ISO 26000 definition of **social responsibility** is as follows: "The responsibility of an organization for the impacts of its decisions and activities on society and the environment, through transparent and ethical behaviour that:

- **contributes to sustainable development**, including health and the welfare of society
- takes into account the expectations of stakeholders
- **is in compliance with applicable law and consistent with international norms of behavior**, such as human rights, the precautionary principle[3] and prevention principle[4], among others;
- and is integrated throughout the organization and practised in its relationships".

This definition implies the following dimensions in assessing societal responsibility:

- some kind of **alignment with the SDGs** defined by the UN (United Nations. Department of Economic and Social Affairs, 2015)
- the importance of considering a **larger community** than the research community,
- the importance of some kind of **organizational structure** to oversee the assessment of decisions and activities according to such principles.

This is in line with the following definition of RRI (Responsible Research and Innovation) by the European Commission: "Responsible Research and Innovation is a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view on the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society)" (European Commission. Directorate-General for Research and innovation & Schomberg, 2011). In other words, it emphasizes that research and innovation have the **responsibility to avoid harm**, **to do good** and to create and support **global governance structures** that can facilitate the two former responsibilities (Buhmann & Fieseler, 2021).

This undescores the point made earlier that social responsibility assessment

[3] "Principle adopted by the UN Conference on thr Environment and Development (1992) that in order to protect the environment, a precautionary approach should be widely applied, meaning that where there are threats of serious or irreversible damage to the environment, lack of full scientific certainty should not be used as a reason for postponing cost-effective measures to prevent environmental degradation. (2) The precautionary principle permits a lower level of proof of harm to be used in policy-making whenever the consequences of waiting for higher levels of proof may be very costly and/or irreversible"(*GEMET - Environmental Thesaurus — European Environment Agency*, 2021)

[4] "This principle allows action to be taken to protect the environment at an early stage. It is now not only a question of repairing damages after they have occurred, but to prevent those damages occurring at all. This principle is not as far-reaching as the precautionary principle. It means in short terms: it is better to prevent than repair" (*GEMET - Environmental Thesaurus — European Environment Agency*, 2021)

- requires **involving stakeholders** from the society in some form,
- is a **continuous process**, that requires discussion and feedback loops at the different stages of an innovation,
- requires **transparency and science outreach** about innovations so that they can be understood, trusted and evaluated by non-specialists and accepted by society.

FD-X30-031 (Association Française de Normalisation, 2013) defines 6 governance action areas[5]:

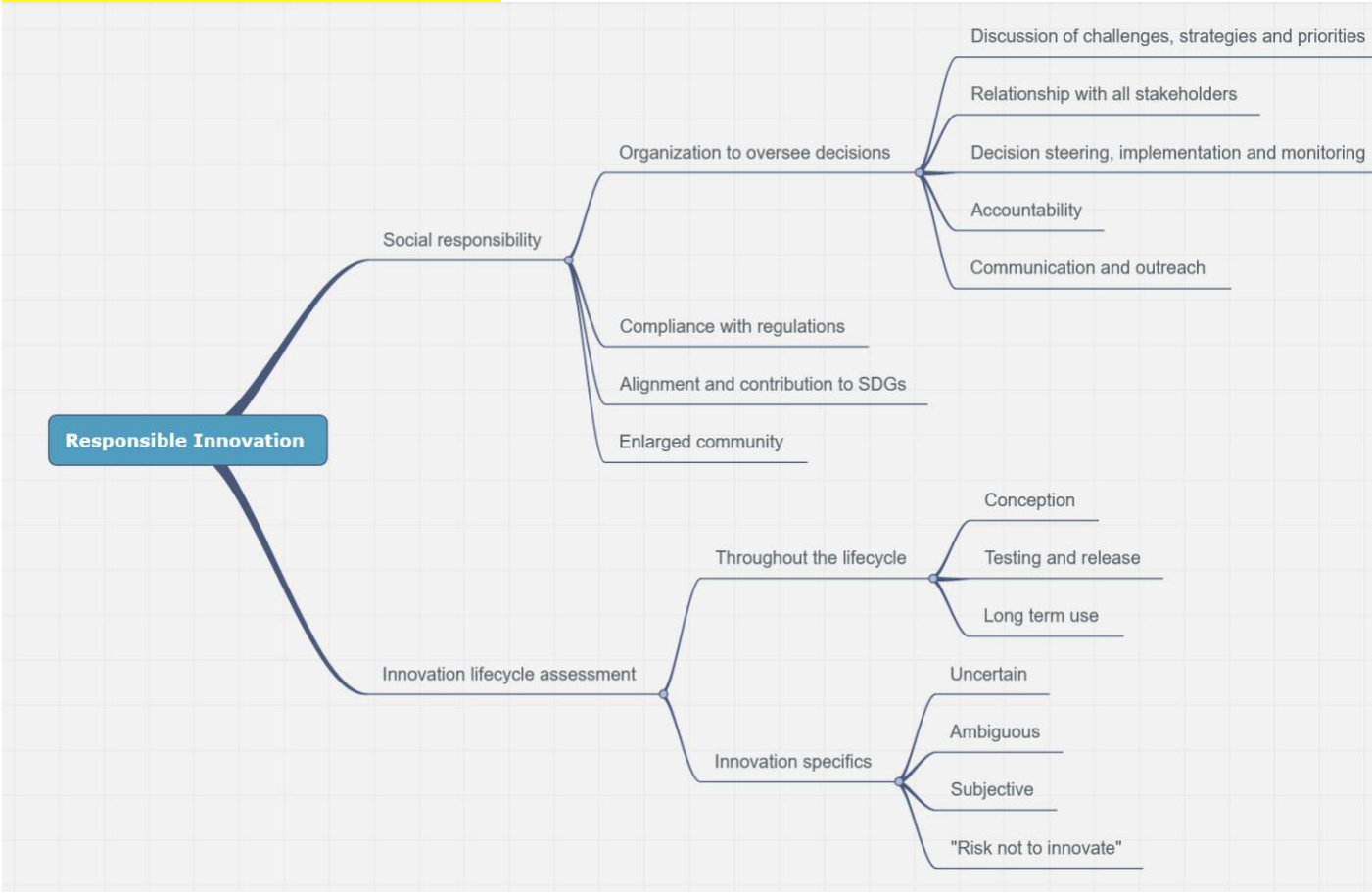
- **responsibility principles, vision, values**, which refers to the organization's missions, ethical aspects, and the representation of the organization in the long term.
- **relationship with stakeholders**, which reiterates the need to involve those impacted by innovation management decisions.
- **analysis of challenges, strategies and priorities**, which stresses the need to find a satisfactory trade-off in an efficient, realistic exchange with stakeholders who, by definition, have limited availability; in other words, the need for formalized processes and transparent, documented and traceable decisions to have a realistic but fair arbitration.
- **structures and decision-making processes**, with the aim of designing them to enable the organization to adapt them to each area of societal responsibility (e.g. each SDG), to represent the diversity of interests and situations, and to promote equitable treatment among stakeholders.
- **steering, implementation and monitoring**, with the aim of controlling that decisions are enforced and to provide guidance for continuous improvement and adaptation of the organization's processes.
- **accountability and communication,** with the aim of regularly communicating decisions to stakeholders and demonstrating that stakeholder input has been taken into account.

Here is an attempt to summarize all the previous concept in a mind map:

[5] Unless explicitly stated, translations from this French norm are from the authors of the present paper.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

Draft awaiting European Commission approval



1. Mind map on the principles of responsible innovation

What concerns should be addressed in priority?

There is no one-size-fits-all solution, as the literature is littered with possible dimensions – the 17 Sustainable Development Goals (SDGs) in themselves show how diverse social responsibility is. Looking at the Special Eurobarometer on Responsible Research and Innovation (Special Eurobarometer 401: Responsible Research and Innovation (RRI), Science and Technology, 2013), there is a hint of what European respondents want from RRI:

> "Respondents[…] want research and innovation to be carried out with due attention to **ethical principles** (76%), **gender balance** (84%), and **public dialogue** (55%). Similar to results of earlier Eurobarometer surveys, more than half of all Europeans are interested in developments in science and technology (53%), but a majority **do not feel informed enough** (58%).

Innovation Management in action: how can we achieve it?

Which organizational solution? Business clusters and Knowledge and Innovation Communities: Think globally, act locally

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

All of the above points to the need for structures that allow for an equitable, ongoing dialogue between innovators (research organizations, private companies) and the potential users and / or impacted stakeholders (civil society). Finding a neutral, trusted organization to facilitate the dialogue is quite challenging, as all stakeholders will have their own interests and biases in assessing Responsible Innovation.[6]

Marcandella suggests the "pôles de compétitivité" which are the French variant of **business clusters**, as good candidates to play such a role (Marcandella, 2015). Business clusters are defined by Porter as "geographic concentrations of interconnected companies, specialized suppliers, service providers, firms in related industries, and associated institutions (e.g., universities, standards agencies, trade associations[7]) in a particular field that compete but also cooperate" (Porter, 2000). We can add the local public powers and tech transfer organizations as other typical members of such an ecosystem. The clusters have several assets in favor of organizing responsible innovation management: among other things,

- they are operating at the **macro-level**, which is a good way to optimize the dialogue with stakeholders on aspects of societal responsibility – but not too broad either, as they are **thematized**;
- they are **local**, which forces dialogues with civil society and local policy makers to be very concrete and operational, and supports accountability to the local population;
- last but not least, they **already bring together all these different stakeholders** and facilitate dialogue between them.

From a top-down perspective, a complementary structure is the **Knowledge and Innovation Communities** (KICs). KICs are consortia emanating from the European Institute of Innovation and Technology (EIT)[8] whose mission since its creation in 2008 is to be a major facilitator of innovation and economic growth in the EU. EITs bring together leading businesses, education and research institutions to form transnational partnerships called KICs. As of 2023 there are nine different KICs (E.U. Funds, 2022):

- **EIT Climate KIC**: Innovation for climate action
- **EIT Digital**: Inclusive, fair and sustainable digital innovation
- **EIT InnoEnergy**: Acceleration of sustainable energy innovations
- **EIT Health**: Boosting innovation in health
- **EIT Raw Materials**: Developing raw materials into a major strength for Europe
- **EIT Food**: Addressing sustainable food supply chains from resources to consumers
- **EIT Manufacturing**: bringing together European manufacturers
- **EIT Urban Mobility**: transforming urban mobility
- **EIT Cultural & Creativity**: Culture and creativity leading sustainable value creation and growth.

KICs are highly complementary to business clusters because of their trans-national nature. They have great potential to bring together local clusters with similar areas of interest and to raise funds.

[6] This does not mean organizations should not integrate social responsibility in their internal decision making processes, but ~~rather than~~ they are not sufficient on achieving socially responsible innovation management on a wider level.

[7] Trade associations being defined by Porter as "competitive assets, not merely lobbying and social organizations". It therefore encompasses economic as well as societal aspects.

[8] https://eufunds.me/what-is-the-eit-european-institute-of-innovation-technology/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

Last but not least, an RRI approach cannot be successful unless there is a real cultural change in the organization to support responsible research and innovation. Griesdoorn and al. report from a recent survey of Dutch policy officers in the specific field of quantum technology that RRI principles are weakly present, if not absent (Griesdoorn et al., 2023). The conclusion sounds like a rule of thumb that can be generalized to other countries and sectors to prevent RRI from being a mere social or greenwashing tool: "This may need more tailormade efforts to fortify RRI principles, for example: courses, training programs or workshops based on the current perspectives of the Dutch policy officers, confronting the Dutch policy officers with their perspectives, and addressing the way RRI and the RRI principles could contribute to specific governmental and societal goals". In other words, organizational sincere commitment is the key of a virtuous RRI, and further, of any assessment tool like an index.

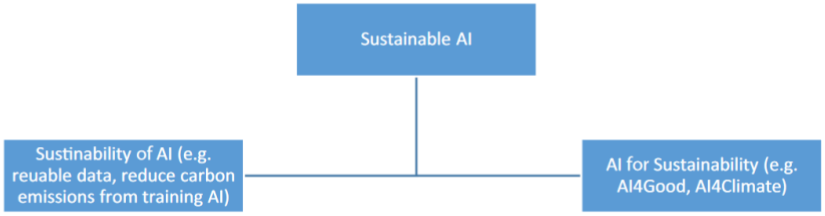
Digital Innovation and AI: what specifics?

The previous sections focused on innovation as a whole. Now we will focus on what is arguably the biggest innovation leap in recent history: Artificial Intelligence. To what extent should it be treated differently from other innovations? What are its core singularities?
In its *White paper on Artificial Intelligence* (European Commission, 2020), the European Commission defines AI as a "collection of technologies that combine data, algorithms and computing power". It highlights that "The use of AI systems can have a significant role in achieving the Sustainable Development Goals, and in supporting the democratic process and social rights […] Given the increasing importance of AI, the environmental impact of AI systems needs to be duly considered throughout their lifecycle and across the entire supply chain, e.g. as regards resource usage for the training of algorithms and the storage of data".

In other words, AI is specific in the sheer scale and variety of contexts in which it can be applied, as its impact is systemic. Moreover, social responsibility can be studied using both external (AI for green) and internal (green AI) criteria, as its utter scale and potentially transformative nature can have a huge impact (positive or negative) on the achievement of the SDGs, but also a huge and increasing environmental impact in terms of the resources it consumes. We can try to illustrate the above concepts and apply them to AI. We have at core the concept of *sustainable AI*, as defined by Van Wynsberghe as "a movement to foster change in the entire lifecycle of AI products (i.e. idea generation, training, re-tuning, implementation, governance) towards greater ecological integrity and social justice" (Van Wynsberghe, 2021), with its dialectic between two concepts:

- AI for sustainability
- Sustainability of AI

Fig. 1 Sustainable AI as sustainability *of* AI vs AI *for* sustainability



2. Fig. 1 cited from Van Wynsberghe, A. (2021). Sustainable AI : AI for sustainability and the sustainability of AI. AI and Ethics, 1(3), 213-218.

First specifics: The multifaceted impacts of AI

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

AI has far-reaching social ramifications and is currently undergoing rapid expansion with an ever-changing state of the art. What's more, the discussion cannot be driven by a single organization, because AI is mostly conceived and designed at the level of a domain, not an organization. This requires communicative and deliberative approaches so that everything can be discussed in an open forum.

Second specifics: The opacity problem

We saw in the previous part that transparent and ongoing dialogue with the stakeholders directly affected by the innovation is a key element of responsible innovation. With AI, however, this is proving to be a serious challenge due to its opacity. There is a *paradigm shift* regarding the nature of opacity when applied to AI: "as machine learning algorithms are not only a set of rules defined by programmers but also contain algorithmically self-produced rules of learning[, t]hese procedures may for practical purposes be structurally inaccessible and incomprehensible not only to laypersons by oftentimes also [...] to the organizations that own and employ them, and even to system programmers and specialists" (Buhmann & Fieseler, 2021). This means that transparence is not enough because the process is inherently opaque: "AI opacity often cannot simply be 'tackled' by demanding that organizations 'make their algorithms transparent' based on a fixed standard or framework". In other words, transparency is necessary (datasets and algorithms used) but not sufficient to achieve "AI literacy" for the wider public (Buhmann & Fieseler, 2021). Buhmann & Fieseler suggest that though AI companies may be unwilling to open up their processes in order to remain competitive, the gain in reputation and trust from the civil society counterbalances this. Civil society's critical gaze and debate is also a great asset for AI, which primarily benefits from and improves upon user feedback.

This leads to the following principles, regarding needs, as stated by Buhman and Fieseler:

- needs to be addressed in an **open forum** where everyone can voice their concerns, opinions and arguments: the debate needs to be public and **multivocal** to reflect the different aspects at stake and the different points of view;
- needs for the participants to have as much **information and understanding** as possible about how the AI works.
- needs the process to be **ongoing and responsive** so that the feedback is taken into account and actually influences recommendations or decisions.

AI and SDGs

AI has great potential to help us address complex, multidisciplinary problems such as those encompassed by the Sustainable Development Goals (SDGs)[9], especially (Vinuesa et al., 2020). In this study, the 17 SDGs and finer-grained corresponding 169 targets were used to assess AI impact. As a result of the study, 79% of the targets are considered to be positively impacted by AI, while 23% of them are considered to be negatively impacted by AI. On some targets, AI can have both a positive and negative impact, depending on the aspect under consideration.

[9] Sustainable Development Goals are a collection of seventeen interlinked objectives designed to serve as a "shared blueprint for peace and prosperity for people and the planet, now and into the future." (United Nations. Department of Economic and Social Affairs, 2015).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

Draft awaiting European Commission approval

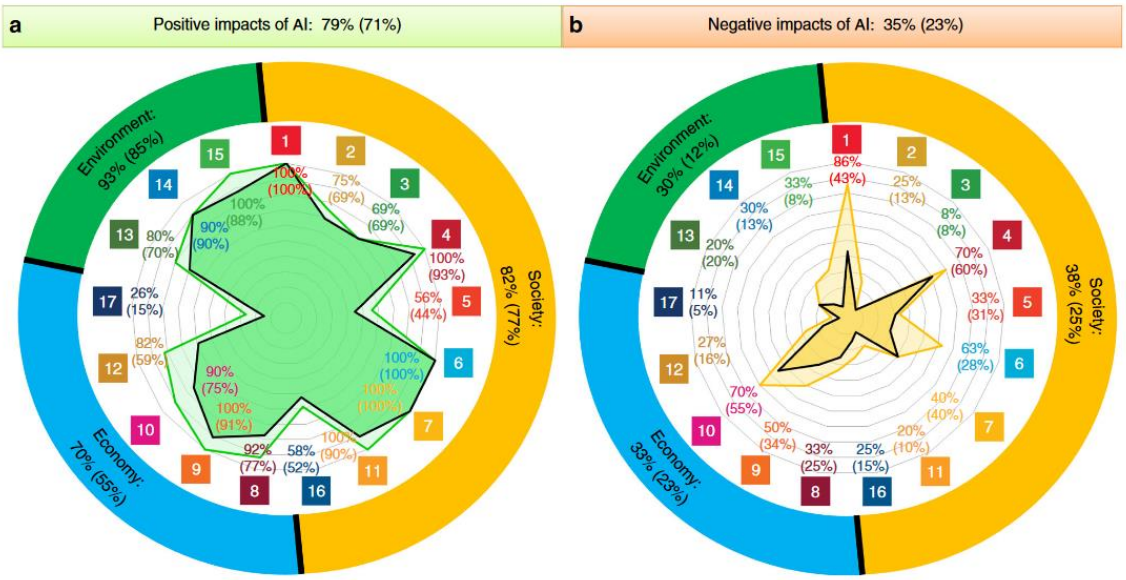


Fig. 1 Summary of positive and negative impact of AI on the various SDGs. Documented evidence of the potential of AI acting as (**a**) an enabler or (**b**) an inhibitor on each of the SDGs. The numbers inside the colored squares represent each of the SDGs (see the Supplementary Data 1). The percentages on the top indicate the proportion of all targets potentially affected by AI and the ones in the inner circle of the figure correspond to proportions within each SDG. The results corresponding to the three main groups, namely Society, Economy, and Environment, are also shown in the outer circle of the figure. The results obtained when the type of evidence is taken into account are shown by the inner shaded area and the values in brackets.

3. Figure cited from Vinuesa et al., 2020.

If we group the SDGs according to the 3 traditional "pillars", this goes as follows:

- Concerning the "**Society**" pillar (SDGs 1-7, 11, 16), 82% targets could potentially benefit from AI, mainly by optimizing workflows: providing food, health, water and energy to the population by enabling smart cities; supporting the circular economy and enhancing energy-efficient systems. However, 31% targets, however, can be negatively impacted. Many of these relate to the implementation and use of AI (of possibility thereof) in countries with different cultural values and wealth. The process of processing big data in AI systems to "nudge" people, known as "big nudging"[10], is completely duplicitous: "Personal data may be used to 'nudge' people to make healthier and environmentally friendly decisions. Yet the same technology may also promote nationalism, fuel hate against minorities or skew election outcomes" (Helbing & Pournaras, 2015). It is a concern for democracy and human rights, as such tools can be used in regions where "ethical scrutiny, transparency, and democratic control are lacking" (Helbing & Pournaras, 2015). Using AI to compute social scores is an example of how AI can lead to an increase in inequalities (as it leads to the differentiation of citizens' rights based on a score). It has also been demonstrated that AI systems can reproduce gender stereotypes because of the source data - see, for example, (Dastin, 2018)) - or due to the gender-biased word embeddings present in the language itself upon which LLM and generative AIs are built (Bolukbasi et al., 2016). Other biases have been

[10] See for instance https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

demonstrated, such as a racial bias in automatic speech recognition systems (Martin & Wright, 2023).

- Concerning the "**Economy**" pillar (SDGs 8-10, 12, 17), 70% of the targets can be positively impacted, and are mostly related to the increased productivity enabled by AI (for instance in supply chain management (Toorajipour et al., 2021)), but 33% of the targets can be negatively impacted. The competitive advantage provided by the AI can lead to an increase in inequality between nations, but also within nations: by replacing old jobs with ones that require more skills, technology disproportionately rewards the educated, resulting in an increase in inequalities (Helbing & Pournaras, 2015) and a "shift of corporate income to those who own company from those who work there" (Vinuesa et al., 2020). The amplification factor of AI-trained recommendation-based algorithms result can lead to political and social polarization that hinders social cohesion. What's more, by stimulating economic growth, AI can have a negative impact on responsible consumption, if we consider that buying things without a need is not aligned with SDG12, even with responsibly produced products.
- For the "**Environment**" pillar (SDGs 13-15), 93% of the targets are positively impacted while 30% can be negatively impacted by AI. This is the area with the highest net impact, but also the one with the most uncertainty. The overall potential of AI, and machine learning in particular, as a tool to help climate change (Rolnick et al., 2023). For example, artificial intelligence can be leveraged to optimize energy consumption and dynamic real-time monitoring of the energy mix, and by its huge ability to process large amounts of data and images can help identify polluted areas such as oil spills (Keramitsoglou et al., 2006) or track desertification trends. The negative impact is the high consumption of energy and natural resources (metals, water) required by AI to function, that are following the explosion of the demand in this area in recent years (see "Green AI" section below). There is also a structural risk in the use of AI: as a potential great tool to acquire extensive, real-time knowledge of ecosystems, it may be lead to the over-exploitation of resources that we were not aware of before (Vinuesa et al., 2020).

Responsible AI and Open Science

Beyond alignment with SDG principles, it should be noted that Open Science and FAIR principles are cornerstones of responsible innovation and thus responsible digital innovation and AI in particular. As responsible innovation means taking into account a world with limited resources, Open Science FAIR principles, where results and source data are made available (but also understandable and reusable), are key for the following reasons:

- They lower the barrier to entry to innovation because results are available to everyone;
- They avoid wasting resources to duplicate work because some findings where unknown and/or unavailable to innovators at a given point in time.

AI is not specific in this regard, and the application of FAIR principles to AI has already been discussed (Huerta & al., 2023). One initiative that adheres to such principles is the training of the BLOOM model[11]: "its architecture, catalogue of data used, and training log are all publicly available, to facilitate research into language models"[12]. What's more, Open Science includes practices such as

[11] See https://huggingface.co/bigscience/bloom.

[12] https://www.cnrs.fr/en/press/release-largest-trained-open-science-multilingual-language-model-ever

citizen and societal engagement at its core, which is very compatible to the concepts of society involvement and debate required in Responsible Research and Innovation.

A focus on education

As an educational tool, AI is essentially ambivalent, especially with the worldwide adoption of generative AI by society at large since the public release of ChatGPT at the end of 2022. Universities and other higher education institutions are gradually positioning themselves on this issue, with different approaches ranging from conservative[13] to permissive, problematized ones (Göttingen, UNIGE).

The potential of AI has been widely discussed in the last couple of years. Mollick and Mollick outlined seven educational approaches to education with AI (Mollick & Mollick, 2023)that we will cite here as a framework to think about AI in education. These seven approaches consist in assigning a specific role to an AI tool in a roleplay game: AI as a mentor, AI as a tutor, AI as a coach, AI as a teammate, AI as a student, AI as simulator, and finally, AI as a... tool.

TABLE 1 SUMMARY OF SEVEN APPROACHES

AI USE	ROLE	PEDAGOGICAL BENEFIT	PEDAGOGICAL RISK
MENTOR	Providing feedback	Frequent feedback improves learning outcomes, even if all advice is not taken.	Not critically examining feedback, which may contain errors.
TUTOR	Direct instruction	Personalized direct instruction is very effective.	Uneven knowledge base of AI. Serious confabulation risks.
COACH	Prompt metacognition	Opportunities for reflection and regulation, which improve learning outcomes.	Tone or style of coaching may not match student. Risks of incorrect advice.
TEAMMATE	Increase team performance	Provide alternate viewpoints, help learning teams function better.	Confabulation and errors. "Personality" conflicts with other team members.
STUDENT	Receive explanations	Teaching others is a powerful learning technique.	Confabulation and argumentation may derail the benefits of teaching.
SIMULATOR	Deliberate practice	Practicing and applying knowledge aids transfer.	Inappropriate fidelity.
TOOL	Accomplish tasks	Helps students accomplish more within the same time frame.	Outsourcing thinking, rather than work.

[4]Table from Mollick and Mollick 2023. Benefits and Risks of using 7 pedagogical approaches using AI

[13] For instance SciencesPo Paris bans the use of ChatGPT without systematic, proper referencing: https://newsroom.sciencespo.fr/sciences-po-bans-the-use-of-chatgpt/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

The table above summarizes the risks and benefits of each approach and gives an indication of the ambivalent nature of AI in education, as it requires critical hindsight and techniques to leverage AI technologies efficiently. In other words, much of the issue boils down to AI literacy, which could be a teaching requirement in higher education whose goal (among others) is to prepare students for their professional lives. As Breit Eika puts it, "Many of our students will have to use artificial intelligence in their professional lives, and we must of course prepare them for this – not only for the technological aspects but also for the legal, ethical and social aspects it involves". In other words, "instead of merely knowing how to use AI applications, learners should be inculcated with the underlying AI concepts for their future career, as well as the ethical concerns of AI applications to become a responsible citizen" (Ng et al., 2021). The following areas around this AI literacy can be summarized as:

- Know and understand AI
- Apply AI
- Evaluate and create AI
- AI ethics.

A focus on AI and education among the Enlight partners

In February 2024, a survey was sent out to Enlight partners (all higher education institutions) concerning how the issue of AI in education was being adressed by the institutions. The survey and compiled answers are available as a separate appendix, as they will quickly become outdated in a rapidly changing field where not all institutions have finalized official public policies but will very most likely do so in a few months. What's more, some analyses (especially on the impact of AI in the institution) are sometimes hampered by a lack of hindsight in a recent, massive change in tools and practices.

The general trends are that no institution has decided to forbid AI in education. The approach of the Enlight Alliance institutions is one of permission, with the majority providing guidance and training to faculty and/or students and adapting teaching and/or assessment methods. In most cases, formal guidance and/or governance documents are either already published or planned. Another important take-home point is that no institution has modulated its approach to AI depending on the educational domain: all guidance and/or regulation initiatives apply to all domains.

"Green AI"?

The environmental impact of AI has been demonstrated and will continue to rise steeply as its adoption and use increase in the coming years. AI, like any other digital service, requires energy and resources to operate. We noted earlier in the paper that innovation should be assessed throughout its lifecycle, and AI is no exception. The following aspects should be assessed (Ligozat et al., 2021; Luccioni, 2023):

- **Data acquisition, production and storage**: systems and energy required to function;
- **Model training**: systems and energy required to operate;
- **Testing and deployment phase**: systems/devices (such as sensors in smart devices or smart buildings) and energy required to build and use them
- In all phases, energy consumption should ideally be evaluated against the carbon intensity of the energy grid used.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

Most Large Language Models do not provide detailed information allowing a detailed Life Cycle Assessment. Such recent initiatives such as the BLOOM model are designed to be transparent, allowing for a meaningful estimation of their carbon footprint (Luccioni, 2023).

A focus on energy consumption

Generative AI is notorius for being energy intensive, which has an environmental impact, among other external negativities of AI. For instance, it has been demonstrated that "training one large NLP [Natural Language Processing] model (aka a transformer), with neural architecture search, resulted in over 600,000 CO2e(lbs), roughly the equivalent of carbon emissions of five cars (over the lifetime of the car)" (Van Wynsberghe, 2021), whilst "GPT-3 needs to "drink" (i.e., consume) a 500ml bottle of water for roughly 10-50 responses, depending on when and where it is deployed" (Li et al., 2023).

How can we limit the energy consumption of AI? Experimentations are underway in several areas:

- Neuromorphic systems (using the brain structure to shape the AI system)
- Memristors (using physics instead of computation for data transmission)
- Bringing sensors and processors closer to one another through edge computing, which requires less data transport, leading to lower power consumption) through event-based sensors
- Spintronic nanodevices where you can have neurons and synapses at the same time: a 100-fold energy gain is expected with these devices.[14]
- The development of Quantum Computing may be a factor of future impact.

However, the first question is to use AI where relevant (is AI efficient for this task?) and to have a reasonably adequate model training, with a trade-off between the accuracy or speed of the AI model and its energy cost: it is not always necessary to have a very fast or very accurate answer, in which case a smaller model is fine and will consume much less energy. For example, in the field of speech-to-text recognition, a "SOTA Transformer emits 50% of its total training released CO2 solely to achieve a final decrease of 0.3 of the word error rate" (Parcollet & Ravanelli, 2021).

Regulatory questions

It goes without saying that responsible DI/AI in Europe should comply with European regulation, with a key text still in progress, the AI Act. Compliance with European regulations is "level zero" for responsible innovation, and also a good indicator of the mindset to be adopted on the matter of responsible AI: indeed, European regulations in the area of digital innovation all have in common the goal of finding the right balance between innovation and economic growth on the one hand, and individual rights and European values on the other.

The GDPR sets the ground rules for what data should be used for what purpose. As far as digital innovation is concerned, in most cases it can be boiled down to (but not limited to) the following aspects:

[14] These areas for optimization have been summarized in the "Green AI" Workshop held in Dec. 2022, see https://www.youtube.com/watch?v=WxZTbnlvnuI in particular the talk from Adrien F. Vincent, "Building Green AI".

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

- **Risk assessment**: Am I collecting data that directly or indirectly relates to individuals? Is it sensitive data[15]?
- **Defined purpose**: Is each atomic data treatment linked to pre-defined goals?
- **User information**: Are individuals informed about what data is collected about them and how it is used? Is it possible to deny?

As of March 2024 - the time of this deliverable -, the AI Act has not been officially adopted yet, however its outline and main principles are stable. Its core approach is based on the level of risk and impact of the AI system, and delineates what is acceptable or not, and under what conditions:

- **Unacceptable risk**: Any AI application in this category is prohibited in the EU. This applies to AI that have the potential to compromise privacy and/or violate fundamental rights as expressed in the EU Charter of Fundamental Rights. This concerns: AI practices that have the potential to manipulate individuals without their conscious knowledge or to exploit vulnerable populations to manipulate their behavior (AI-driven subliminal techniques); AI-based social scoring systems; real-time biometric identification systems in public spaces.
- **High risk**: These are targeted at sensitive sectors such as health, education, recruitment, credit scoring, law enforcement or justice; and/or components of products already subject to EU safety regulations (such as toys, vehicles, medical devices[16] and lifts). These require "full, effective and properly documented *ex ante* compliance conformance with all requirements of the regulation and compliance with robust quality and risk management systems and post-market monitoring". This requires high quality data for training and transparency about the data used to prevent bias, but also the commitment to perform post-market bias monitoring, detection and correction, and public availability of the technical documentation. Common European Data Spaces[17] are explicitly mentioned as instrumental for high quality data for AI training.
- **Limited risk**: These encompass the AI systems that require transparency on their automated nature: either because of their ambiguity (chatbots that interact with humans or automatically generated content, such as deep fakes); or because they are using "near-subject" data, such as emotion detection or biometric data, to make decisions. In those cases, transparency is required about the automated nature of the process and what ~~the~~ data it uses.
- **Minimal or no risk**: the AI systems that do not fall into the previous three categories and do not have a direct impact on privacy or transparency issues. Tools such as spam filters, AI used in video games and others fall into this category. They are not subject to specific restrictions.

All of the previous discussions on responsible AI suggest that

- **Regulation**/risk mitigation and **cybersecurity** measures are key whenever there is heavy reliance on AI for automated decisions; risk mitigation and cybersecurity measures are needed to avoid increasing vulnerability

[15] Sensitive data consists of "[so-called] racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation". Cf. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32016R0679

[16] Medical devices are also regulated in the 2017/745 EU Medical Device Regulation, with the question of how both regulations work together because the AI Act requires a conformance work and the Medical Device Regulation requires a certification process.

[17] https://digital-strategy.ec.europa.eu/en/policies/data-spaces

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

- Responsible AI requires some level of **transparency** about the data used for training and the methods used to fine-tune it, as well as some level of **scientific outreach** to explain how it works, its limitations and potential benefits. This question is all the more important because AI is now in a situation where it is both widely adopted, with the explosion of public use of generative AIs, and used as a black box.
- AI applications seem particularly good at ***optimizing*** **resources and processes**. This raises a red flag about a potential Jevons effect that can lead to over-exploitation of resources to achieve maximum efficiency.
- **AI consumes energy and resources** at all stages of its lifecycle. All other things being equal, AI services that are transparent about their design and carbon impact should be preferred and fostered, and model choice and training tailored to their required response time and/or accuracy.
- Responsible AI means using it in an **organization/ecosystem that implements *Responsible Innovation Management***, in particular open and regular assessment and debate in a public forum. To be effective and balanced, this requires enlightening civil society about how AI works: AI literacy is a key concept in which higher education institutions have a crucial role to play in the future.

Conclusion: towards a DI/AI index

How can we build indicators to assess the "responsible" nature of DI/AI?
The state of the art in the domain of responsible DI/AI, summarized in this document, points out some do's and don'ts:

Innovation as an ongoing process: avoid the "one shot" effect (an assessment of the DI/AI in the organization once a year, or an assessment of an innovation just after its release). The index should be designed to encourage assessment throughout the lifecycle of an innovation.

Criticality assessment: European regulations in the digital domain often define categories of risk that lead to different risk mitigation measures (and, in most risky scenarios, prohibit them). The AI Act, which has not been released yet, adopts an approach by classifying the potential external negativities of an AI into 4 risk categories, whilst the GDPR has differentiated processes depending on the criticity of the data (personal data, sensitive data). The DI/AI Index should take this into account as there is a necessary tension between innovation and research and their responsible design or use; what's more the Jevons effect is very important to take into account in the environmental and social impact of an innovation, and it can only be measured *after* an innovation has been released and used by civil society.

Indicator definition and testing: indicators should be carefully tested to ensure that they actually lead to a reliable, rational notation. In particular, the index should avoid at all costs a design, or process, that allows it to be used for green or social washing.

Purpose of the index: the index could be used as a guide for decision making, but also as a checklist for innovation stakeholders to easily measure the impact of their innovation and take measures to mitigate negative externalities. In the long run, such an index could also be used as a policy tool, which leads to the aspect below.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

Index as a starting point, not an end: it should be noted that the index cannot have a significant impact if it is used as a stand-alone tool: it requires institutional policy and/or enforcement measures to be effective. This aspect should have an important part of the assessment. This means that it should include not only the necessary technical elements of measure (e.g., carbon footprint, energy consumption), but also the organizational aspects: is the society involved in the evaluation process of an innovation, with sufficient knowledge to provide informed feedback? Is the matter of responsible innovation integrated into the organization's processes?

We hope the report allows one to see very clearly that the responsible innovation is not a technical issue, but rather a societal one, with key organizational and educational aspects so that there is an open and balanced dialogue, debate and assessment among informed actors. An index should be viewed and designed as a tool to help achieve this wider goal.

Bibliography

Association Française de Normalisation. (2013). *FD-X30-031: Responsabilité sociétale—Gouvernance et responsabilité sociétale—ISO 26000*.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings* (arXiv:1607.06520). arXiv. https://doi.org/10.48550/arXiv.1607.06520

Buhmann, A., & Fieseler, C. (2021). Towards a deliberative framework for responsible innovation in artificial intelligence. *Technology in Society*, *64*, 101475. https://doi.org/10.1016/j.techsoc.2020.101475

Collingridge dilemma. (2023). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Collingridge_dilemma&oldid=1151189090

Dastin, J. (2018, October 11). Insight—Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/idUSKCN1MK0AG/

E.U. Funds. (2022, May 6). *What is the EIT (European Institute of Innovation & Technology)?* EU Funds. https://eufunds.me/what-is-the-eit-european-institute-of-innovation-technology/

European Commission. (2020). *White Paper on Artificial Intelligence A European approach to excellence and trust*.

European Commission. Directorate-General for Research and innovation, & Schomberg, R. von. (2011). *Towards responsible research and innovation in the information and communication technologies and security technologies fields*. Publications Office. https://data.europa.eu/doi/10.2777/58723

Fernez-Walch, S., & Romon, F. (2016). *Management de l'innovation: De la stratégie aux projets* (4e édition.). Vuibert.

GEMET - Environmental thesaurus—European Environment Agency. (2021). [Glossary]. https://www.eionet.europa.eu/gemet/en/themes/

Griesdoorn, F., Kroesen, M., Vermaas, P., & Van De Poel, I. (2023). The presence of Responsible Research and Innovation in the perspectives of Dutch policy officers regarding innovation with quantum technology. *Journal of Responsible Technology*, *16*, 100071. https://doi.org/10.1016/j.jrt.2023.100071

Helbing, D., & Pournaras, E. (2015). Society: Build digital democracy. *Nature (London)*, *527*(7576), 33–34. https://doi.org/10.1038/527033a

International Organization for Standardization. (2018). *Innovation management—Innovation management system—Guidance (ISO Standard No. 56002:2019)*.

International Organization for Standardization. (2020). *Guidance on social responsibility (ISO Standard No. 26000:2020)*.

Kelly, B. G., & Habib, M. (2023). Missed period? The significance of period-tracking applications in a post-Roe America. *Sexual and Reproductive Health Matters*, *31*(4), 2238940. https://doi.org/10.1080/26410397.2023.2238940

Keramitsoglou, I., Cartalis, C., & Kiranoudis, C. T. (2006). Automatic identification of oil spills on satellite images. *Environmental Modelling & Software : With Environment Data News*, *21*(5), 640–652. https://doi.org/10.1016/j.envsoft.2004.11.010

Ligozat, A.-L., Lefèvre, J., Bugeau, A., & Combaz, J. (2021). Unraveling the Hidden Environmental Impacts of AI Solutions for Environment. *arXiv.Org*. https://doi.org/10.48550/arxiv.2110.11822

Luccioni, D. S. (2023). *Towards Measuring and Mitigating the Environmental Impacts of Large Language Models*. CIFAR. https://cifar.ca/cifarnews/2023/09/25/cifar-ai-insights-towards-measuring-and-mitigating-the-environmental-impacts-of-large-language-models/

Marcandella, E. (2015). Management responsable de l'innovation-Concept, méthodologie, perspectives. *QUALITA'2015*. https://hal.science/hal-01149772/

Martin, J. L., & Wright, K. E. (2023). Bias in Automatic Speech Recognition: The Case of African American Language. *Applied Linguistics*, *44*(4), 613–630. https://doi.org/10.1093/applin/amac066

Mollick, E., & Mollick, L. (2023). *Assigning AI: Seven Approaches for Students, with Prompts* (arXiv:2306.10052). arXiv. https://doi.org/10.48550/arXiv.2306.10052

Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021). AI Literacy: Definition, Teaching, Evaluation and Ethical Issues. *Proceedings of the Association for Information Science and Technology*, *58*(1), 504–509. https://doi.org/10.1002/pra2.487

Pavie, X. (2012). *Innovation-responsable: Stratégie et levier de croissance des organisations*. Eyrolles.

Porter, M. E. (2000). Location, Competition, and Economic Development: Local Clusters in a Global Economy. *Economic Development Quarterly*, *14*(1), 15–34. https://doi.org/10.1177/089124240001400105

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A. S., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C. P., Ng, A. Y., Hassabis, D., Platt, J. C., … Bengio, Y. (2023). Tackling Climate Change with Machine Learning. *ACM Computing Surveys*, *55*(2), 1–96. https://doi.org/10.1145/3485128

Sauzet, F. (2022). *Du projet innovant au management responsable de l'innovation: Créez un produit dont le monde a vraiment besoin*. AFNOR.

Silberzahn, P. (2020, July 6). *Tech for good: Et si c'était une très mauvaise idée?* Philippe Silberzahn. https://philippesilberzahn.com/2020/07/06/tech-for-good-et-si-mauvaise-question/

Special Eurobarometer 401: Responsible Research and Innovation (RRI), Science and Technology. (2013). European Commission. https://europa.eu/eurobarometer/api/deliverable/download/file?deliverableId=40745

Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P., & Fischl, M. (2021). Artificial intelligence in supply chain management: A systematic literature review. *Journal of Business Research*, *122*, 502–517. https://doi.org/10.1016/j.jbusres.2020.09.009

United Nations. Department of Economic and Social Affairs. (2015). *The 17 goals | Sustainable Development*. https://sdgs.un.org/goals

Draft awaiting European Commission approval

Van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, *1*(3), 213–218. https://doi.org/10.1007/s43681-021-00043-6

Vinuesa, R., Azizpour, H., Leite, I., Balaam, M., Dignum, V., Domisch, S., Felländer, A., Langhans, S. D., Tegmark, M., & Fuso Nerini, F. (2020). The role of artificial intelligence in achieving the Sustainable Development Goals. *Nature Communications*, *11*(1). Scopus. https://doi.org/10.1038/s41467-019-14108-y

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

Draft awaiting European Commission approval

Appendix A: Survey on AI in education

Content of the survey

What is your institution?

Regarding the matter of AI in education, does your institution...

- ☐ forbid the use of generative AIs for education
- ☐ allow the use of AIs for education
- ☐ allow the use of AIs for education but adapts its training and evaluation methods
- ☐ provide trainings to help students and teachers use generative AIs like ChatGPT or the like
- ☐ Other

Other :

Does your institution provide documents that offer guidance and/or regulate the use of AI in education*?

- ○ Yes, for guidance and regulation
- ○ Yes, only for guidance
- ○ Yes, only for regulation
- ○ No
- ○ It is currently under development

**This survey is focused on education and is therefore limited to pre-doctorate studies (first and second cycle, e.g. bachelor's and master's degrees)*

Please provide a link to this document:

What kind of impact did the AI guidance and/or regulations have in your institution?

- ☐ Curriculum development: integration of AI-related questions (ethics, basic features and limits...) to existing curricula
- ☐ Change in teaching practices
- ☐ Change in assessment and evaluation practices
- ☐ Creation of a unit dedicated to "AI in education" (task force or other) in your institution
- ☐ Other

Other :

In which domains does your institution provide such guidance and/or regulation?

- ☐ All domains
- ☐ Social Sciences and Humanities
- ☐ Physical Sciences and Engineering
- ☐ Life Sciences
- ☐ Other

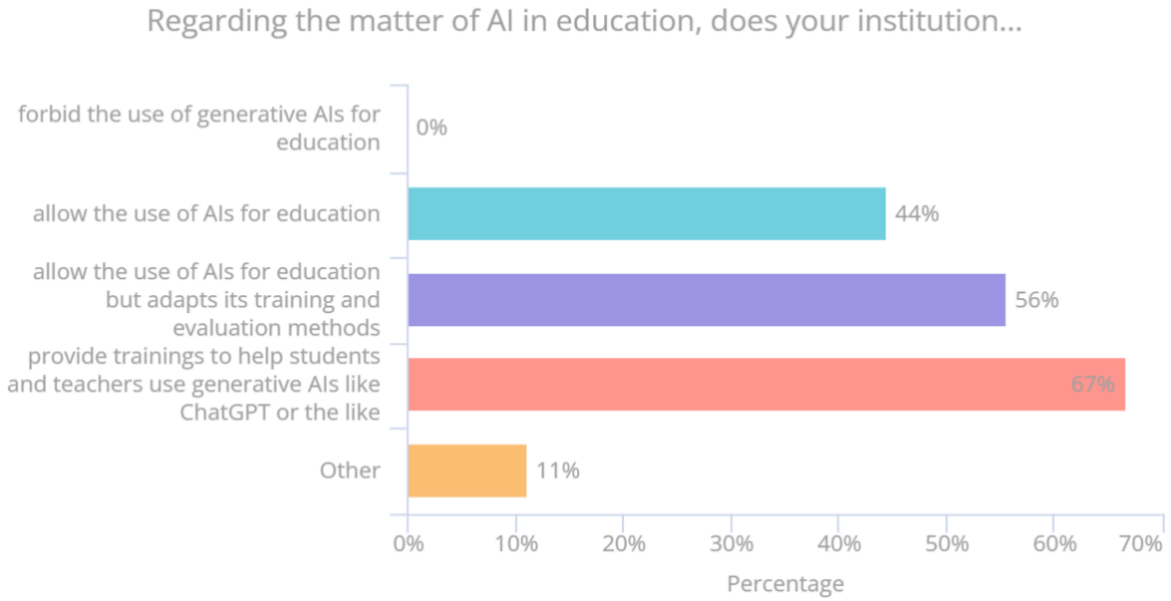
Other :

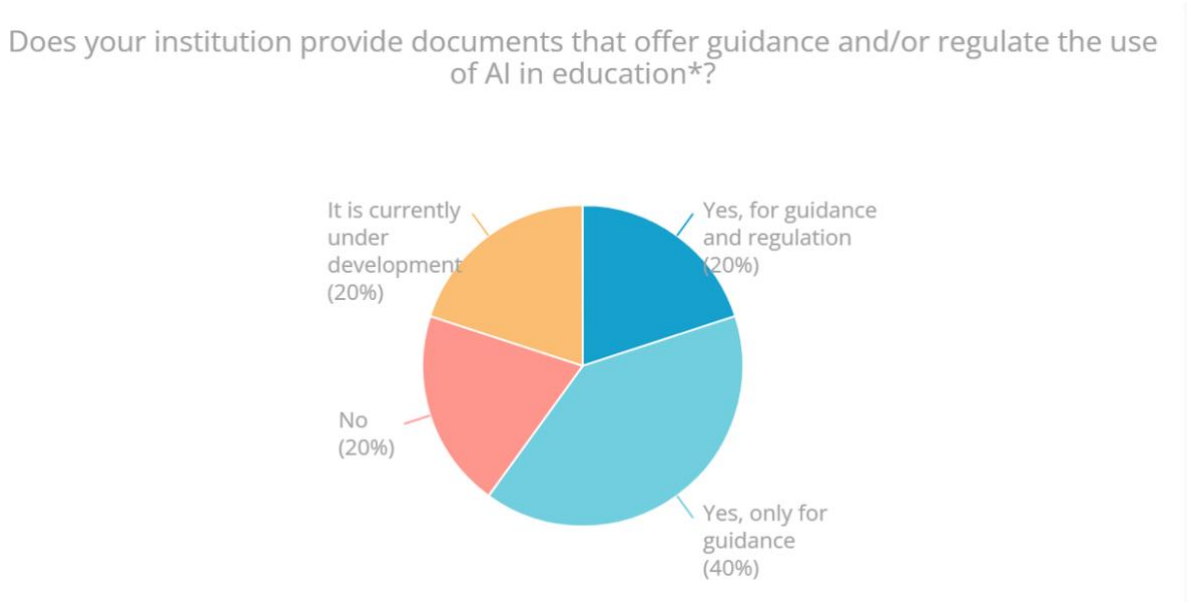
Compiled results

All the partners answered with the exception of Bern University.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

Draft awaiting European Commission approval



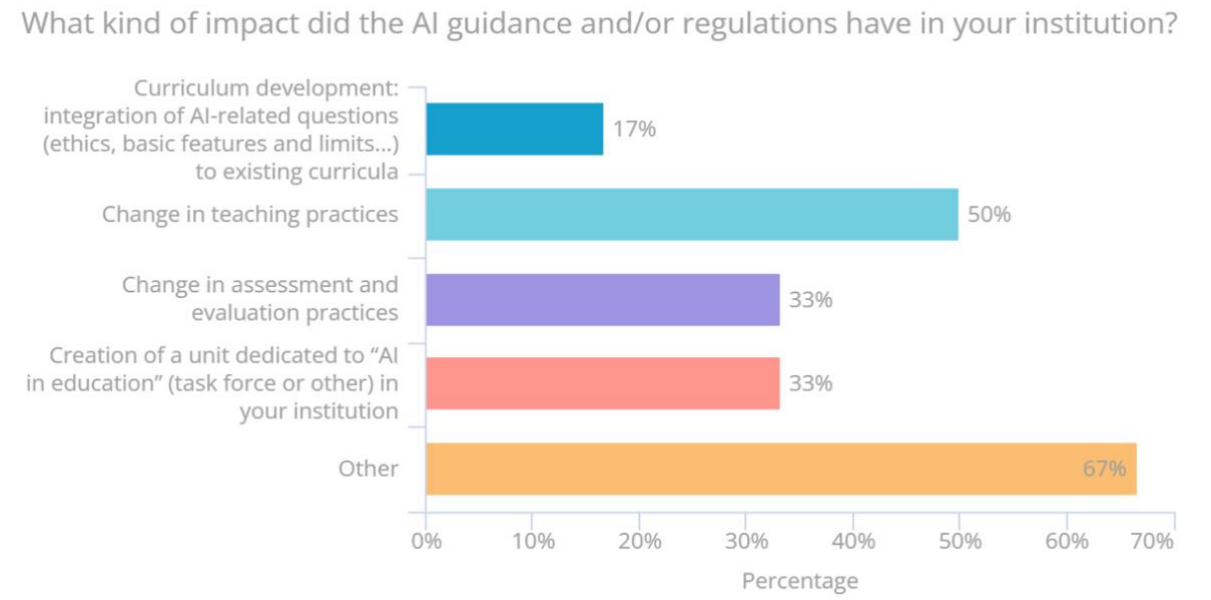


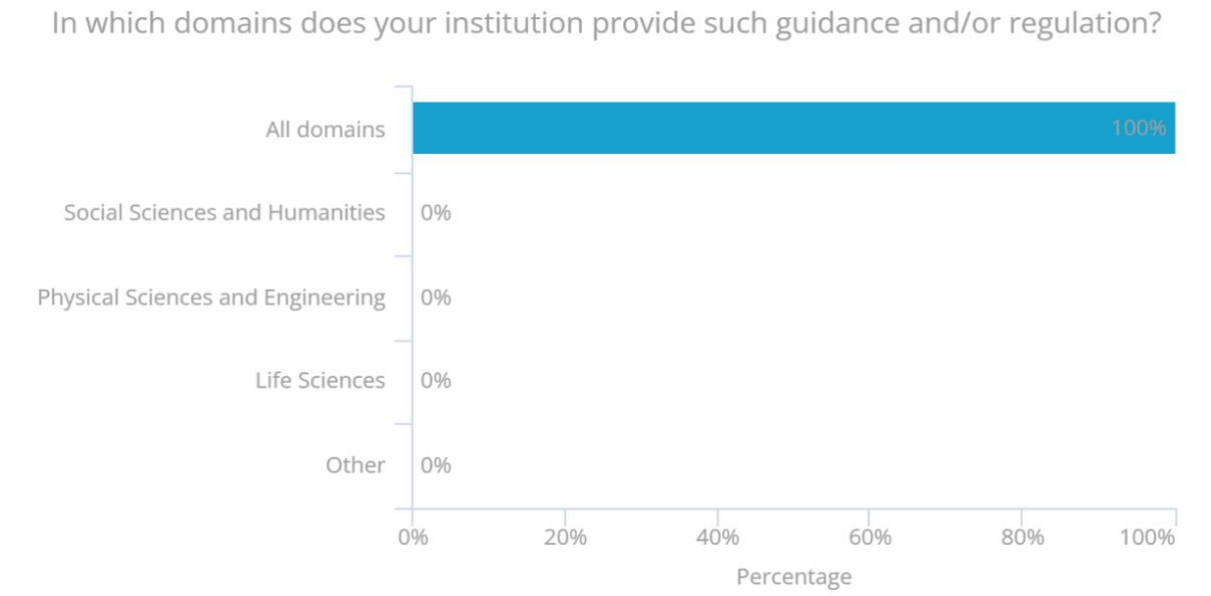
Nature of the documents:

- Public documents (Tartu, Göttingen, Groeningen, Bordeaux)
- Public documents at national level (Galway)
- Internal documents (Ghent)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

Draft awaiting European Commission approval





Public documents from partner institutions on institutional policy / guidance on AI in education:

- University of Bordeaux: https://enseigner.u-bordeaux.fr/outils-et-ressources/IAG
- University of Göttingen: https://uni-goettingen.de/en/674738.html
- University of Galway: https://www.qqi.ie/sites/default/files/2023-09/NAIN%20Generative%20AI%20Guidelines%20for%20Educators%202023.pdf (national guidelines from the higher education sector)
- University of Groeningen: https://edusupport.rug.nl/2365784080 and https://www.rug.nl/about-ug/organization/service-departments/teaching-academy-groningen/communities-of-practice/a-i-in-education
- University of Tartu: https://ut.ee/et/sisu/university-tartu-guidelines-using-ai-chatbots-teaching-and-studies

Appendix B: Tentative summary of pros and cons´ of AI per SDG

The table below was an attempt to generate an automated classification with prompts entered ChatGPT[18], then edited. Internal aspects ("greenness" of AI as a technology) are added with the mention "internal". Due to the initial bias and in spite of the editing, this result should be taken with a grain of salt.

SDG	AI possible positive effects	AI possible negative effects
Poverty Alleviation (1)	Create economic opportunities, automate routine tasks, and enhance productivity. Example: AI-powered job matching platforms	AI can lead to job displacement or deletion, particularly in industries where automation is prevalent, potentially exacerbating unemployment and poverty.
Zero Hunger (2)	AI can optimize agricultural practices, improve crop yields, and enhance food distribution systems to address food security challenges. AI applications for precision agriculture, like autonomous drones and sensors, can monitor and manage crop health, leading to increased agricultural productivity and reduced food waste.	The use of AI in agriculture can be inaccessible to small-scale farmers due to cost and technological barriers, potentially widening the gap between large and small agricultural enterprises.
Good Health and Well-being (3)	AI can revolutionize healthcare by providing personalized medicine, early disease detection, and improving healthcare delivery systems. Example: AI-based diagnostic tools, such as machine learning algorithms analyzing medical images, can enhance early detection of diseases, improving treatment outcomes and reducing mortality rates.	Privacy concerns and potential misuse of personal health data in AI applications can hinder the adoption of AI in healthcare and compromise patient trust.
Quality Education (4)	AI can enhance access to education, personalize learning experiences, and bridge gaps in educational resources.	AI-driven personalized learning can reinforce educational inequalities if access to technology and digital resources is unevenly distributed among students.
Gender Equality (5)	AI technologies could be developed and used in ways that avoid reinforcing gender biases and	AI systems can perpetuate gender biases present in training data, leading to unfair outcomes and reinforcing existing inequalities.

[18] Model used: GPT 3.5, prompts entered: "how can you align artificial intelligence with the SDGs? please provide arguments and examples for each item"; "can you do the same but adding 1 counter argument and 1 counter example ?" (January 11th, 2024)

	contribute to closing gender gaps in various sectors.	
Clean Water and Sanitation (6)	AI can assist in optimizing water management systems, predicting water quality issues, and ensuring efficient water distribution.	Dependence on AI for water management can lead to vulnerabilities in critical infrastructure, raising concerns about system reliability and potential cyber threats. **Internal**: AI training requires a massive amount of energy and computing power, which uses water for buildings and for use.
Affordable and Clean Energy (7)	AI can optimize energy production and consumption, enhance energy efficiency, and support the transition to renewable energy sources	The production and disposal of AI hardware contributes to electronic waste and environmental degradation, counteracting the goal of sustainable energy solutions. **Internal**: The energy sources used by AI may not by itself be renewable.
Decent Work and Economic Growth (8)	AI can stimulate economic growth by fostering innovation, improving productivity, and creating new job opportunities in emerging industries	Concentration of AI development in a few tech hubs leads to geographical disparities in job opportunities, contributing to regional economic imbalances. **Internal**: The fine-tuning of general-purpose AI can be based on digital labor.
Industry, Innovation, and Infrastructure (9)	AI can drive innovation, improve infrastructure planning, and contribute to the development of sustainable technologies	Rapid advancements in AI may outpace the ability of regulatory frameworks to address ethical concerns, potentially leading to unintended consequences.
Reduced Inequality (10)	AI can be used to identify and address inequalities by providing data-driven insights and creating more inclusive systems	AI technologies can exacerbate existing social inequalities if access to, knowledge of, and control over AI resources is concentrated in the hands of a few powerful entities.
Sustainable Cities and Communities (11)	AI can contribute to the development of smart and sustainable urban environments by optimizing resource use and improving services	Smart city initiatives relying heavily on AI can face resistance from citizens concerned about privacy and surveillance and may be subject to potential cyber threats, leading to increased vulnerability.
Responsible Consumption and Production (12)	AI can help optimize resource use, reduce waste, and promote sustainable consumption patterns	**Internal:** The energy consumption of AI models and infrastructure may contribute to increased carbon emissions, contradicting efforts towards sustainable production and consumption.
Climate Action (13)	AI can assist in monitoring climate change, predicting environmental	**Internal:** The environmental impact of manufacturing and maintaining AI hardware may outweigh the benefits of

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101035819

	trends, and developing solutions for mitigation and adaptation	using AI in climate modeling and mitigation efforts.
Life Below Water (14) and Life on Land (15)	AI can be used to monitor and protect biodiversity, prevent illegal activities, and promote sustainable management of oceans and ecosystems	Overreliance on AI for conservation efforts may lead to a neglect of traditional ecological knowledge and community-based conservation practices.
Peace, Justice, and Strong Institutions (16)	AI can enhance the efficiency of legal systems, support crime prevention, and contribute to building transparent and accountable institutions	Biases in AI algorithms used in legal systems may result in unfair and discriminatory outcomes, undermining the goal of achieving justice for all.
Partnerships for the Goals (17)	Collaboration and partnerships are essential for harnessing the full potential of AI in achieving the SDGs, involving governments, businesses, academia, and civil society.	Unequal power dynamics in global AI partnerships may lead to the exploitation of developing countries and reinforce existing economic disparities.