Trustworthy AI

Framework for Trustworthy AI education V1.0

September 2021

Table of Contents

Intellectual output holders: Andrea Aler Tubella, Juan Carlos Nieves, Umeå University, Sweden

Interviews conducted by:
Andrea Aler Tubella, Umeå University, Sweden
Marçal Mora-Cantallops, Universidad de Alcalá, Spain
Douglas Cirqueira, Maynooth University, Ireland
Noah Schöppl, ALLAI, Netherlands
Christofer Talvitie, ALLAI, Netherlands

Template: Orla Casey, Momentum

Revision of the document: all Trustworthy AI partners

1. Overview

This document has been developed in the context of the Erasmus+ project "Trustworthy AI" with the goal of facilitating the introduction of the High-Level Expert Group's Guidelines on Trustworthy AI (EUCommission, 2018) into Higher Education across disciplines. This goal is aligned with the EU's digital strategy, which emphasises the need to train professionals that can "shape technology in a way that respects European values" (EU, 2020). Towards this purpose, the High-Level Expert Group's Guidelines on Trustworthy AI outline the necessary requirements for responsible and trustworthy development; the goal of this project is therefore to use them as a starting block for the introduction of ethical and socio-legal competences in Higher Education topics related to AI.

The main goal of this Framework is to describe the principles and learning strategies to be followed to develop students' competences. The findings are presented in the form of **recommendations for educators**, **needs in educational materials** and **policy incentives** answering the following questions:

1. What strategies are needed for effectively introducing the High-Level Expert Group's requirements in Higher Education?
2. Which competences and learning outcomes related to Trustworthy AI should Higher Education students develop? How to assess them?
3. Which resources can facilitate the introduction of Trustworthy AI education?
4. What risks and opportunities are presented by the introduction of Trustworthy AI education? Which policy incentives can facilitate the opportunities?

The Framework we present is based on the **analysis of 11 expert interviews across 5 countries** as well as a study of the literature in the form of a **systematic literature review**.

Throughout this document, we will refer to the High-Level Expert Group as "HLEG", to the Guidelines on Trustworthy AI as "the Guidelines", to the associated Assessment List as "the Assessment List" and to the 7 requirements outlined in the list as "the Requirements". Higher Education will sometimes be abbreviated as "HE".

2. Methods

2.1. Literature review

A key focus of the Guidelines is to incorporate ethical and socio-legal perspectives into the development and use of AI. However, laws, social norms and ethical dimensions are highly contextual (Turiel, 2001). Thus, in order to understand the state of the art across higher education, a systematic literature review (Appendix C) was undertaken with the aim to **analyse the development and assessment of ethical competences in Higher Education** in a variety of fields. The main questions for analysis are:

1. Which competences should HE students develop to gain proficiency in ethical and socio-legal aspects?
2. How to teach and evaluate these competencies in the context of HE?

We conducted the literature search on Scopus, to obtain results in a variety of subject areas. Papers were identified by the following search terms appearing in title, abstract or key: "*ethics AND teaching AND "higher education" AND (competence OR competency OR skills)".* We limited our search to publications from 2015 onwards, and retrieved a total of 50 publications. We focused on individual research output, so excluded 1 book, 1 editorial, 1 extended abstract and 6 review articles (either paper reviews or curricula reviews). Finally, we manually excluded 13 papers whose topic did not touch upon teaching student skills related to ethics (most of the discarded papers addressed the ethics of teaching, competences for teachers, or were not in the scope of HE). Finally, 4 papers could not be retrieved, leaving a total of 24 papers for analysis. The papers analyse cover a wide variety of subject areas. Based on author affiliation 17 countries are covered, as well as 12 subject areas as indexed by Scopus.
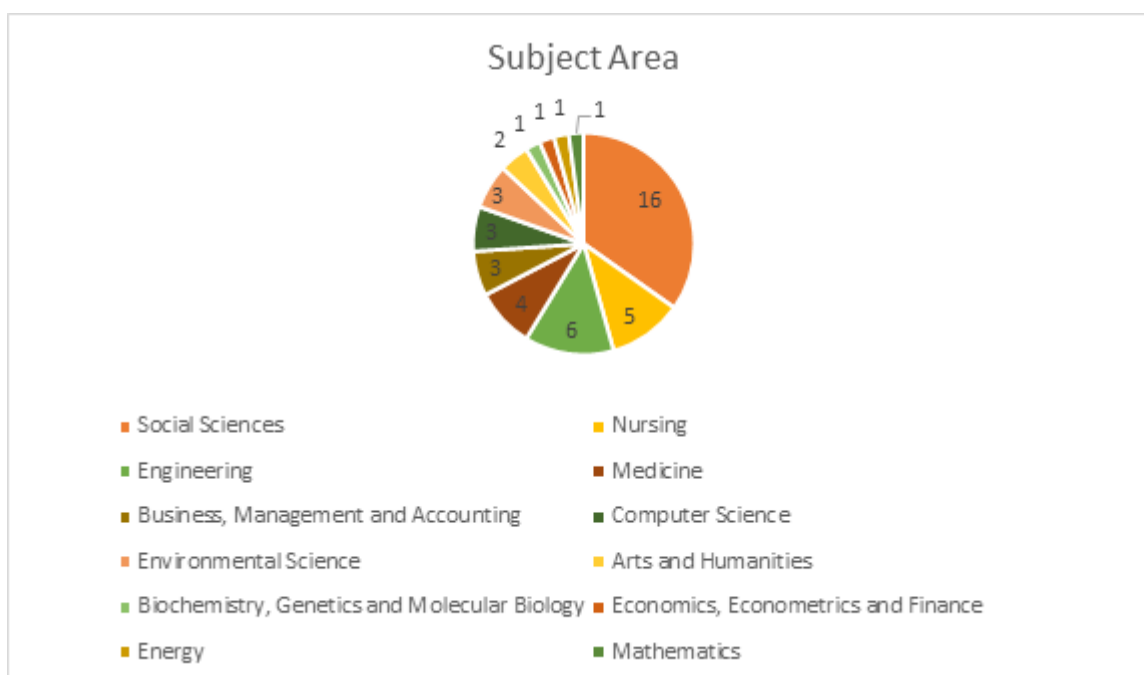
Trustworthy **AI**

2.2. Expert Interviews

With the goal of exploring the state-of-the-art of Trustworthy AI in Higher Education, Umeå Universitet developed an interview protocol (see Appendix B). The specific goals of this protocol were to **obtain expert feedback** on the following topics:

1. General awareness of the Guidelines amongst stakeholders in HE.
2. Inclusion of the Requirements in current educational programs.
3. Current educational practices for Trustworthy AI (topics, learning outcomes, evaluation).
4. Incentives to facilitate the inclusion of Trustworthy AI topics in HE
5. Risks and opportunities.

Partners from ALLAI, Universidad de Alcalá, Maynooth University and Umeå Universitet followed a training session in order to unify how the interviews were conducted. Interviewees were therefore asked the same questions in the same manner, allowing to contrast answers in a qualitative analysis.

A total of 11 interviewees were selected for their involvement in HE, whether through governance, program management or teaching. Interviews were conducted over a period of 6 weeks. The experts, with affiliations in 5 different countries, brought use cases spanning medicine, law, computer science and social sciences (see Appendix A for a breakdown of participants). The responses from expert interviewees inform the recommendations made in this framework and shed light on the current state of Trustworthy AI in education.

3.Requirements

Transforming the assessment list of the ethical guidelines into specific skills for the actors involved in AI development has been highlighted by the European Commission as a natural step in creating an "ecosystem of trust" for the flourishing of European AI (EU, 2020). Thus, the first focus of this framework is an in-depth dive on the requirements of the assessment list. We first analyse how the HLEG requirements are currently included in education, and then turn our focus to individual requirements to understand their perceived importance and the key educational questions surrounding them. From this analysis, we identify several recommendations for effectively involving the assessment list and the HLEG requirements in Higher Education.

Trustworthy **AI**

3.1. State of the art

Are the HLEG requirements currently included in any form in the course/program?
10 responses
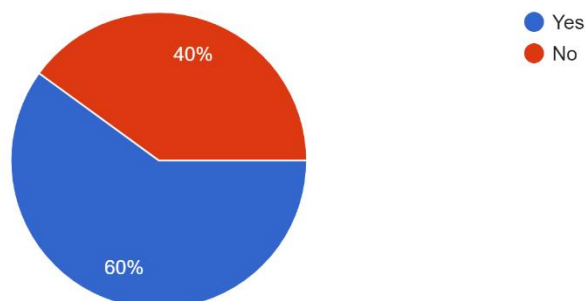
In terms of the inclusion of the HLEG requirements in current courses and programs, **60% of the experts interviewed state that they are currently included in their education case**. A common thread in the expert interviews (7 out of 11) is that **while different requirements are certainly covered in education, they are not explicitly related to AI or to the HLEG guidelines**. Many report that topics around trustworthiness are tackled because of their relevance, but that the relationship to the HLEG requirements is often implicit rather than a deliberate effort.

The HLEG Guidelines and Assessment list are valued for the clarity they bring in specifying different requirements. The documents as a whole may not be suitable for direct inclusion in courses, but the focus on the humans behind the system is fundamental and should always be kept.

The guidelines and assessment list in their current form are valued by all interviewees for setting down clear requirements and bringing clarity to their meaning. However, their inclusion in education raises some challenges: respondents raise that the length and technical nature of the documents is **not suitable for all disciplines and education levels**. In addition, a perspective on how each requirement applies to different disciplines is missing. Furthermore, the challenge of the translation of different technical terms may bring different perspectives depending on which language of the guidelines is being studied. A frequent point made by the experts is that **different courses may touch upon only a few of the requirements**, therefore not looking at the guidelines as a whole, but rather focusing on a few specific relevant requirements. Overall, several respondents note that the key aspect of the Assessment List is **the focus on the human behind the system** and emphasise the value of conveying to students that the responsibility and ethical obligations of AI development lies on those involved in the process.

Trustworthy **AI**

Summary: Which of the HLEG requirements are already being taught in their education program?
10 responses

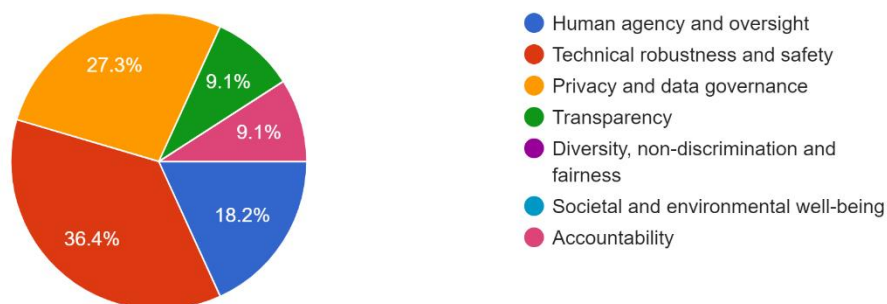| Requirement	Responses
Human agency and oversight	6 (60%)
Technical robustness and safety	5 (50%)
Privacy and data governance	9 (90%)
Transparency	6 (60%)
Diversity, non-discrimination an…	7 (70%)
Societal and environmental wel…	4 (40%)
Accountability	4 (40%)
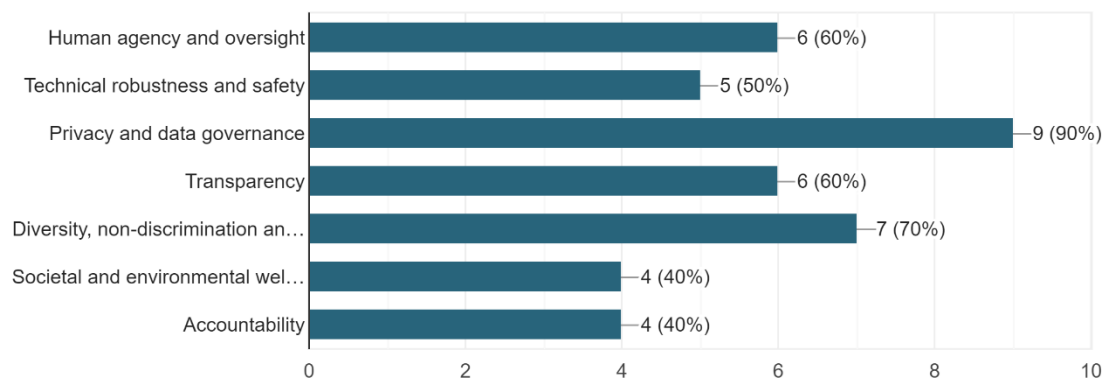
Although the HLEG requirements are currently included in education programs as they intersect with other topics, they are not often addressed through the focus of the Guidelines.

When looking at requirements individually, a clear stand-out in terms of its prevalence in current education programs is **"Privacy and Data Governance", currently included in 90% of the education programs brought up by the interviewees**. Following the idea of implicitness, interviewees state that this requirement is tackled through the lens of good data practices, GDPR or ethics, rather than as a specific focus on the requirement itself. On the other end of the spectrum, **"Societal and environmental well-being" and "Accountability" are the least included**, featuring in 40% of the interviewees' programs.

3.2. Individual requirements and their importance in education

A key theme in the expert interviews is an assessment of the requirements in terms of their importance in education. In particular, interviewees were asked to rank requirements in order of their relevance for their education case. There is significant consensus among the experts (100% of respondents) that **all requirements are relevant**, but that their **significance and importance for a course varies depending on the topic and the area.** For this reason, there was no significant agreement in the ordering of requirements: each education case elicited different orderings depending on the application area and topics tackled in the course.

Trustworthy **AI**

Highest ranked requirement

11 responses

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental well-being
- Accountability

Lowest ranked requirement

11 responses

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental well-being
- Accountability
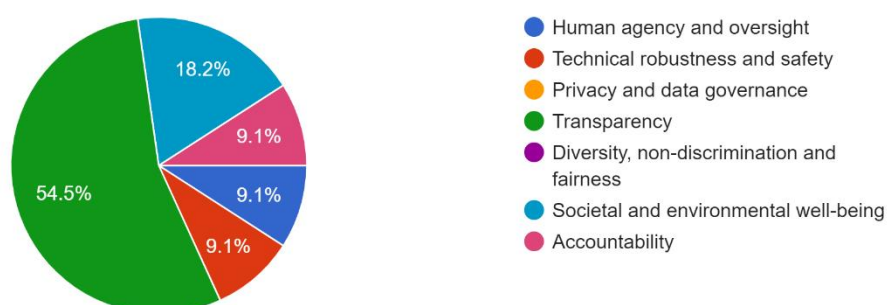
The consensus is that all requirements are relevant for education. The relative importance of each requirement is highly dependent on the topic and application area of a course.

Despite the big variation in rankings, "Transparency" stands out as being ranked the least relevant in 54.5% of the education cases. The reasons for this rating are however very disparate: some experts believe transparency is encompassed by other requirements, while others believe other requirements use more basic concepts that are easier for beginners, and others see the rest of the requirements as more fundamental.

2.3. Recommendations

The state of the art indicates that current education does not focus on the HLEG Guidelines, but rather that different aspects of Trustworthy AI are tackled if and when they intersect with relevant course content. This finding points to a gap in HE, both for educators and students, in terms of awareness of the guidelines and their content. In this sense, it is clear that a two-pronged approach in resources is needed, targeting both the educators and the students.

Firstly, there is a **need for training** and availability of materials for educators of different disciplines to become aware of and gain knowledge on Trustworthy AI from the perspective of the guidelines. This need is echoed by many interviewees (5 out of 11), who mention that a barrier for incorporating Trustworthy AI education may be a lack of experience or guidance for the educators, either because of the novelty of the topic, its interdisciplinarity, or a lack of time to get familiar with these aspects.

Secondly, many interviewees feel like the expertise to teach about different requirements is spread across disciplines and topics. In this sense, rather than having a single course focused on the Guidelines, most consider it more natural to include teaching on the Requirements in already existing courses where they are relevant. Although no single requirement is seen as more valuable, many interviewers report that each requirement may or may not be relevant depending on the course and the discipline. In this sense, **clarifying the importance of each requirement for different course contents** is a necessary step to make the inclusion of the Guidelines into education explicit.

Overall, the assessment of the HLEG Requirements in terms of their importance in education yields the following recommendations for educators:

> **R1.** Explicitly include HLEG Requirements in **existing courses** when relevant.
>
> **R2. Bridge the gap between course content and HLEG Requirements** by being explicit about which requirements are being tackled in the course and how.

4.Learning outcomes and assessment

4.1. Learning outcomes

When assessing competencies related to incorporating ethical and social dimensions into Higher Education in all disciplines, results of our literature review indicate an emphasis on *dual* competence (Kim Brown et al., 2019; Noah & Aziz, 2020; Trobec & Starcic, 2015; Zamora-Polo & Sánchez-Martín, 2019): developing technical competence *alongside* the ability to understand and act according to ethical and social expectations. Although discussion on *specific* learning outcomes is notably absent, three learning goals are prevalent for demonstrating mastery of social and ethical competencies (see Appendix C):

- Ethical appreciation/sensitivity: Identifying and understanding the ethical and moral dimensions of a situation.
- Ethical analysis: Deliberating about actions, how they relate to ethical guidelines and codes of conduct, and their possible consequences.
- Ethical decision-making/Applied ethics: Selecting and implementing a course of action in response to ethical reasoning.

These findings squarely align with syllabi analysis, where it has been found that the most common sought outcomes for teaching Tech Ethics are variations on "recognize/critique/reason" (Fiesler, Garrett, & Beard, 2020).

Thus, the following learning outcomes emerge for each individual requirement when adapting these three levels of competence:

LO1. Appreciation: Identifying the applicability of the requirement in different contexts and its different dimensions for different stakeholders.

LO2. Analysis: Deliberating about possible implementations of the requirement, how they relate to ethical guidelines and codes of conduct, and their possible consequences.

LO3. Application: Selecting and technically implementing a solution in response to analysis in terms of the requirement.

The combination of these learning objectives would indicate a high level of mastery for students at three different edges: identification, reasoning and implementation.

4.2. Learning outcomes per requirement

Turning our focus to specific learning objectives for each requirement, the interviews shed light on how they are currently being addressed in education. Interviewees were asked to provide example of how they address certain requirements in their educational case, as well as which questions around each requirement they deem more valuable for HE. A breakdown per requirement can be summarised by:

- **Human agency and oversight**
  The questions raised around this requirement revolve around 3 edges: **appropriate human control, legal frameworks, and inclusiveness.** Firstly, recognising and implementing the appropriate level of human control to guarantee human agency and oversight. Secondly, understanding the legal constraints of the application area in terms of decision making (e.g. in the medical domain). Lastly, taking a strong human rights perspective and guaranteeing inclusivity for equal agency and oversight across society.

- **Technical robustness and safety**
  Interviewees call for education around this requirement to be specific to AI, with **accuracy and reliability** at the forefront, with the questions of how to recognise them and ensure them. Further, some respondents stress the importance of teaching to **balance technical robustness and ethical constraints**.

- **Accountability**
  Questions around this requirement revolve around **auditing and record keeping, legal frameworks for liability, and demonstrating minimisation of negative effects**. This discussion includes a focus on how to audit intelligent systems and how to ensure traceability so that systems can be effectively audited. In addition, there is a call for teaching students in technical areas about legal aspects related to liability and demonstration of due diligence in mitigating negative effects.

- **Privacy and data governance**
  Questions raised on "Privacy and data governance" all surround data, specifically **its quality, provenance and privacy preservation**. The education questions mentioned are how to collect and recognise quality data, maintain privacy and prevent biases in the data and the models built from it. In addition, there is a focus on the legal requirements for data collection and storage (e.g. GDPR).

- **Transparency**
  Education questions for "Transparency" are very multifaceted, but there is significant consensus among respondents. A first question involves educating students on **recognising transparent systems** and how they differ from opaque ones. Further, many interviewees relate transparency to explainability and raise the question of educating students on being able to explain system's decisions or reasoning and **providing the skills to develop explainable AI**. An additional question relates transparency to traceability: educating developers on how to **document and expose the data, processes and decisions taken in the design process**. Finally, there is a last topic revolving around **transparency of data processing**: being explicit about how data is collected and what is done with it.

- **Societal and environmental well-being; Diversity, non-discrimination and fairness**
  Questions around these requirements have significant overlap, addressing the topic of interdisciplinary expertise to address the difficulty of defining or quantifying well-being or diversity. Additionally, some experts point to idea of learning to conduct or read impact assessments to evaluate the effects of systems.

Overall, the questions raised on each requirement, although specific to each, have two common threads: recognition and implementation. On one hand, there is a strong call for teaching students **how to recognise whether a requirement is being followed**. On the other hand, many questions revolve around technical **methods for trustworthy AI development,** e.g. record-keeping methods, privacy preserving data collection methods, explainability methods.

Questions on individual requirements reveal two parallel competence needs: recognition and implementation.

Additionally, interviewees were asked how these requirements were addressed currently in their education case. Interestingly, Trustworthy AI concepts are often introduced in technical courses, coming up as a result of a direct application of the technology on real-life cases. This suggests that ongoing societal concerns play a big role in what is taught, as **educators find it easier to bridge the gap on Trustworthy AI concepts when they can relate them to well-known issues**.

4.3. Assessment

Across the literature review and the expert interviews, there is a noticeable lack of consistent methodologies for assessing soft competences such as ethical and social awareness or understanding and application of guidelines and codes of conduct. The assessment methods uncovered in the literature review mostly rely on self-assessment – e.g. (Ibáñez-Carrasco, Worthington, Rourke, & Hastings, 2020; Mulot-Bausière et al., 2016) – or experts' perception of student's knowledge without explicit grading criteria – e.g. (DeSimone, 2019; Lapuzina, Lisachuk, & Romanov, 2018). On the other hand, interviewees either report no explicit assessment of competences related to Trustworthy AI or include it as part of the overall assessment of programming projects.

The notable lack of structured approaches to assessment constitutes a barrier both for assessing student skills and for measuring the effectiveness of different teaching practices. **Pedagogical guidelines on how to conduct such an assessment are therefore greatly needed.**

4.4. Recommendations

Both the literature analysis and the expert interviews reveal the need for two different levels of expertise. The first is the call for educating on how to **recognise whether a requirement is being followed**, and how. This competence corresponds to LO1 as identified in the literature review: understanding what a requirement *means* in the context of a certain application. In fact, this type of question universally applies to students as citizens, as it allows for identifying and adopting trustworthy technology. In addition, it provides an initial maturity level in terms of understanding the HLEG Requirements.

The second competence identified across requirements corresponds to **technical methods for trustworthy AI development.** There is consensus across interviewees about the need to teach concrete methods for explainability, traceability, data collection, impact assessment, etc. This necessity closely relates to LO2 and LO3 as identified in the literature: knowledge of the available technical tools is necessary to be able to make an informed choice and implement it. Since the relevant techniques vary greatly depending on the topic and area of the course, it is particularly important to explicitly include in the curriculum which topics and methods will be addressed (Bates et al., 2020).

Finally, it is worth observing that a popular way to introduce Trustworthy AI concepts in the classroom is to discuss current social concerns with the applications of the technology studied in the course. In fact, 6 out of 11 interviewees believe that it would be valuable to **relate the abstract requirements set up by the guidelines to more practical terms**- either through real-world examples, industry participation or concrete tools to experiment with different concepts in class.

The analysis leads to the following recommendations in terms of learning outcomes:

R3. Set out clear learning outcomes that describe the level of proficiency expected from the student. Recommended learning outcomes for each individual requirement are:

> **LO1. Appreciation:** Identifying the applicability of the requirement in different contexts and its different dimensions for different stakeholders.
>
> **LO2. Analysis:** Deliberating about possible implementations of the requirement, how they relate to ethical guidelines and codes of conduct, and their possible consequences.
>
> **LO3. Application:** Selecting and technically implementing a solution in response to analysis in terms of the requirement.

R4. Explicitly include Trustworthy AI **development methodologies** in curricula (e.g. record keeping procedures, privacy-preserving data collection methods, explainability tools).

R5. Relate the Requirements to practical knowledge taught in the course. Examples of topics per requirement include:

Human agency and oversight	Appropriate human control, legal frameworks, stakeholder identification and participation
Technical robustness and safety	Accuracy, reliability, balancing technical and non-technical requirements
Accountability	Auditing, record-keeping, liability frameworks, minimisation of harm
Privacy and data governance	Data quality, data collection, privacy methodologies, GDPR compliance
Transparency	Process transparency, explainability
Societal and environmental well-being; Diversity, non-discrimination and fairness	Interdisciplinary collaboration, impact assessment

5.Educational Resources

5.1. State of the art

Uniformly across interviews, experts mention that they **do not use any specific resources related to Trustworthy AI**. Rather, some mention the use of current topical examples, case studies, and relevant literature. This lack of resources is emphasised in several of the interviews by an added emphasis on the lack of training and time to get acquainted with Trustworthy AI and find available teaching resources.

Similarly to the interview results, the literature review reveals that ethical and moral reasoning skills are often taught through **student-led methods** focused on encouraging reflection and debate amongst students: case studies (Lapuzina et al., 2018), role playing (Trobec & Starcic, 2015), debate (K. Brown et al., 2019), experiential

learning (Ibáñez-Carrasco et al., 2020). This observation aligns with findings from other literature reviews, which emphasise the prevalence of games, role playing and case studies in Engineering and Computer Science Education (Hoffmann & Cross, 2021). Although wide-spread, there is however dissent in the literature, where some advocate for more formal training in e.g. moral philosophy, in contrast to student-led activities (Aközer & Aközer, 2017).

The teaching strategies most often used to teach Trustworthy AI aspects influence the type of resources currently available. Indeed, diverse bodies have developed openly available **case studies** on AI ethics, such as Princeton University[1], Santa Clara University[2], University of Washington[3] and UNESCO[4].

5.2. Needs

When interviewees were asked about what type of resources would be useful for integrating Trustworthy AI in Higher Education, several themes emerged. Firstly, 5 out of 11 interviewees coincided in asking for **use cases**. Interestingly, there was significant consensus on the type of use cases deemed necessary: they should be *realistic* and *implementable*. Indeed, using real cases brought directly from the industry that mimic situations where graduating students may find themselves in is seen as important for the usefulness of these scenarios. In contrast with the literature, where use cases are often used for reflection and debate, several interviewees suggested that use cases should be used for practical exploration, where they can implement and "play with" different solutions.

Another frequent mention is a need for material to aid in evaluation, i.e. **exercises or assignments** with a grading guide that can be directly used for assessing students. Indeed, several interviewees shared the difficulty of evaluating knowledge of abstract concepts. A final shared theme was the need for **resources for teachers.** This need was particularly emphasised in relation to the interdisciplinarity needed to breach all aspects of the requirements. Thus, interviewees requested resources that could help them learn about different requirements and understand how to fit them into their curriculum. These included introductory material on the Guidelines, syllabi or evaluation guidelines that educators could use to inform the design of their own courses. In addition, to ensure a multi-sided view on the Guidelines, a suggestion was to provide guidance on of what kind of experts could provide insight on the different topics, or provide recorded lectures on different aspects of the Requirements from scholars in different fields and the industry.

This analysis leads to the following **identified needs**:

> **N1.** There is a need for **real use cases** where students can use practical knowledge in a hands-on manner.
>
> **N2.** There is a need for **exercises and assignments** related to each requirement, together **with grading schemes** for their assessment.
>
> **N3.** There is a need for **training material for educators** to inform the addition of material about the Requirements in their own courses. These may include introductory material on the Guidelines, syllabi, evaluation guidelines, guides on what kind of invited speakers are relevant, or recorded lectures by experts in different disciplines.

6. Strategies, incentives, risks and opportunities

## 6.1. Strategies and incentives

Uniformly, 6 out of 11 interviewees state that they are **not aware of any specific policy strategies** to include aspects of Trustworthy AI into education, either at the level of their institution or at a national level. Simultaneously, 5 out of

11 interviewees mention that the topic of Trustworthy AI is **gaining importance in their organisation**, and that they are actively considering how to include it in their programs. This mismatch indicates that even though Trustworthy AI is being introduced into HE, the effort is mainly driven by the educators themselves rather than by organisational or national strategies. This approach presents the risk of a mismatch in competences between programs in different HE institutions, as the introduction of Trustworthy AI into educational programs is carried out independently rather than within a coordinated strategy.

When asked about when Trustworthy AI should be introduced in education, respondents were split between school level (primary or secondary education) and undergraduate level. In any case, this indicates a consensus that these aspects should not be relegated to specialised degrees at the postgraduate level, but rather should be introduced at an earlier more general competence level.

Summary: In which education cycle should trustworthy AI education need to start?
11 responses
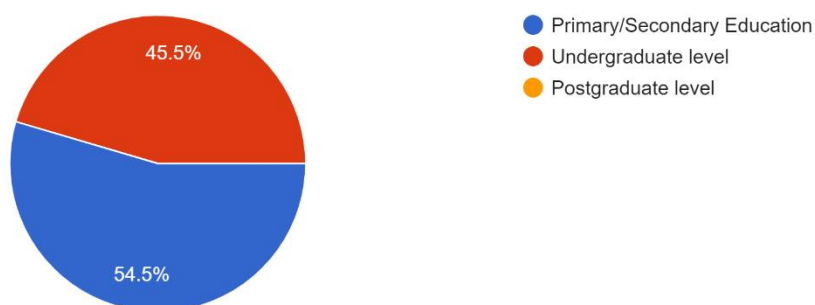
In terms of policy needs and incentives to boost the introduction of Trustworthy AI in HE, interviewees delivered a big variety of suggestions. A big point of consensus (5 out of 11 interviewees) is the **need for investing in Trustworthy AI expertise** so that educators are equipped to teach these topics: this can take the form of investing into multidisciplinary training or boosting the hiring of experts in Trustworthy AI aspects to participate in education. This idea aligns with interviewees mentions of lack of time to get acquainted with the topics in order to be prepared to introduce them in the classroom.

Several interviewees (3 out of 11) mention the importance of allowing for flexibility in the degree structure to allow for the inclusion of broader interdisciplinary topics. They mention that current policies strictly constrain the learning goals of different programs and leave little room for interdepartmental collaboration and interdisciplinarity. In contrast, Trustworthy AI is seen as a topic that would benefit from student's exposure to different disciplines, calling **for policy incentives that will encourage interdisciplinary learning.** These thoughts align with recent calls for transversal education that allows for interdisciplinarity when considering ethics in technology (Raji, Scheuerman, & Amironesei, 2021).

Another factor mentioned is the involvement of the industry, both in providing topical use cases and in playing a role in the employability of students with knowledge of Trustworthy AI. In addition, the role of academic funding was also mentioned, calling for funding agencies to consider Trustworthy AI aspects before granting funding.

6.2. Risks and opportunities

When asked about risks, there was significant consensus amongst interviewees (6 out of 11) in mentioning that there is the risk of introducing Trustworthy AI in HE before institutions are able to prepare, i.e. before there is enough expertise in the topic to be able to teach it competently. In addition, another relevant risk mentioned is that it is important that students from *all disciplines* should be able to learn about Trustworthy AI. Whereas it seems that it is starting to be a focus in STEM, there were some concerns that other disciplines may not be exposed to the topic in HE.

In terms of opportunities, many interviewees (5 out of 11) emphasised that aspects of Trustworthy AI are important for students not only as future professionals, but also as citizens. In this sense, they emphasised the benefits of training a generation of professionals that will possess interdisciplinary knowledge and be able to communicate with professionals from other disciplines on the terms of Trustworthy AI.

6.3. Recommendations

In view of the consensus expressed in the interviews, we identify the following policy recommendations to incentivise the introduction of Trustworthy AI in HE curricula:

> **P1.** Coordinate the introduction of Trustworthy AI in curricula **through national education strategies**, ensuring a uniform adoption.
>
> **P2.** Incentivise HE institutions to **obtain the relevant expertise** needed to teach Trustworthy AI, both by investing resources in training for educators and by hiring experts.
>
> **P3.** Incentivise **interdisciplinary collaboration in education** by valuing it in the curriculum and introducing credits for it.

7. References

Aközer, M., & Aközer, E. (2017). Ethics Teaching in Higher Education for Principled Reasoning: A Gateway for Reconciling Scientific Practice with Ethical Deliberation. Science and Engineering Ethics, 23(3), 825–860. https://doi.org/10.1007/s11948-016-9813-y

Bates, J., Cameron, D., Checco, A., Clough, P., Hopfgartner, F., Mazumdar, S., … de la Vega de León, A. (2020). Integrating fate/critical data studies into data science curricula: Where are we going and how do we get there? In FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 425–435). https://doi.org/10.1145/3351095.3372832

Brown, K., Connelly, S., Lovelock, B., Mainvil, L., Mather, D., Roberts, H., … Shephard, K. (2019). Do we teach our students to share and to care? Research in Post-Compulsory Education, 24(4), 462–481. https://doi.org/10.1080/13596748.2019.1654693

Brown, Kim, Connelly, S., Lovelock, B., Mainvil, L., Mather, D., Roberts, H., … Shephard, K. (2019). Do we teach our students to share and to care? Research in Post-Compulsory Education, 24(4), 462–481. https://doi.org/10.1080/13596748.2019.1654693

DeSimone, B. B. (2019). Curriculum Redesign to Build the Moral Courage Values of Accelerated Bachelor's Degree Nursing Students. SAGE Open Nursing, 5. https://doi.org/10.1177/2377960819827086

EU. (2020). WHITE PAPER On Artificial Intelligence - A European approach to excellence and trust).

EUCommission. (2018). ETHICS GUIDELINES FOR TRUSTWORTHY AI. High-Level Expert Group on Artificial Intelligence, (December). Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Fiesler, C., Garrett, N., & Beard, N. (2020). What Do We Teach When We Teach Tech Ethics? A Syllabi Analysis. https://doi.org/10.1145/3328778.3366825

Hoffmann, A. L., & Cross, K. A. (2021). Teaching Data Ethics: Foundations and Possibilities from Engineering and Computer Science Ethics Education. University of Washington Research Works Archiveo. APA. Retrieved from http://hdl.handle.net/1773/46921.

Ibáñez-Carrasco, F., Worthington, C., Rourke, S., & Hastings, C. (2020). Universities without walls: A blended delivery approach to training the next generation of hiv researchers in canada. International Journal of Environmental Research and Public Health, 17(12), 1–12. https://doi.org/10.3390/ijerph17124265

Lapuzina, O., Lisachuk, L., & Romanov, Y. (2018). Professional Ethics as an Important Part of Engineer Training in Technical Higher Education Institutions. The New Educational Review, 54(4), 110–121. https://doi.org/10.15804/tner.2018.54.4.09

Mulot-Bausière, M., Gallé-Gaudin, C., Montaz, L., Burucoa, B., Mallet, D., & Denis-Delpierre, N. (2016). Formation des internes en médecine de la douleur et médecine palliative : bilan et suggestions des étudiants. Médecine Palliative : Soins de Support - Accompagnement - Éthique, 15(3), 143–150. https://doi.org/10.1016/j.medpal.2015.12.002

Noah, J. B., & Aziz, A. B. A. (2020). A case study on the development of soft skills among TESL graduates in a University. Universal Journal of Educational Research, 8(10), 4610–4617. https://doi.org/10.13189/ujer.2020.081029

Raji, I. D., Scheuerman, M. K., & Amironesei, R. (2021). "you can't sit with us": Exclusionary pedagogy in AI ethics education. In FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 515–525). New York, NY, USA: ACM. https://doi.org/10.1145/3442188.3445914

Trobec, I., & Starcic, A. I. (2015). Developing nursing ethical competences online versus in the traditional classroom. Nursing Ethics, 22(3), 352–366. https://doi.org/10.1177/0969733014533241

Turiel, E. (2001). The Culture of Morality. The Culture of Morality. Cambridge University Press. https://doi.org/10.1017/cbo9780511613500

Zamora-Polo, F., & Sánchez-Martín, J. (2019). Teaching for a better world. Sustainability and Sustainable Development Goals in the construction of a change-maker university. Sustainability (Switzerland), 11(15). https://doi.org/10.3390/su11154224

Appendix A.

Interview profiles

Appendix A. Interview profiles

A total of 11 interviewees were selected for their involvement in HE, whether through governance, program management or teaching.

This appendix contains an anonymised breakdown of their profiles and the use cases they brought to ground the interview on a specific topic.

Table 1: Interviewees' Geographical location as given by affiliation (note that some interviewees have several affiliations).

Organisational affiliation	Number of interviewees
Croatia	1
Ireland	3
Netherlands	2
Spain	3
Sweden	2
EU	1

Table 2: Interviewees' profile.

Position	Number of interviewees
Program coordinator	4
Professor	4
Lecturer	4
Policy maker	1

Table 3: Disciplines of the use cases selected by the interviewees.

Discipline of the use case	Number of interviewees
Computer Science	6
Medicine	1
Law	1
Social Sciences	1
Media Studies	1
Multidiscipline	1

Appendix B.
Interview Protocol

Appendix B. Practical Interview Protocol – Erasmus Plus Project

Juan Carlos Nieves, Andrea Aler Tubella
Department of Computing Science, Umeå University
Email – juan.carlos.nieves@umu.se, andrea.aler@umu.se

Abstract

With the goal of exploring the state-of-the-art of Trustworthy AI in Higher Education, we developed this interview protocol. The specific goals of this protocol were to obtain expert feedback on the following topics: 1. General awareness of the Guidelines amongst stakeholders in HE 2. Inclusion of the Requirements in current educational programs 3. Current educational practices for Trustworthy AI (topics, learning outcomes, evaluation) 4. Incentives to facilitate the inclusion of Trustworthy AI topics in HE 5. Risks and opportunities Partners from ALLAI, Universidad de Alcalá, Maynooth University and Umeå Universitet followed a training session in order to unify how the interviews were conducted. Interviewees were therefore asked the same questions in the same manner, allowing to contrast answers in a qualitative analysis.

Purpose of the Protocol

This protocol is meant as a guideline for interviewers. For the purposes of this document, "you" can refer to the reader (if read within a statement) or the interviewee (if read within a question).

What would be the best-case scenario? In the best-case scenario, this project helps to improve and redesign education programs in Higher Education (HE) in the field/scope of Trustworthy AI.

As a side-effect, the interviews may help the participating entities may get a better view of the aims of the Trustworthy AI guidelines and to reflect on how these affect or are aligned with their own views and education programs.

Our goal is to receive feedback about the following aspects:
1. What is the understanding of HLEG guidelines?
2. How useful is the assessment list for Trustworthy AI in education?
3. How relevant is it for Higher Education?
4. How clear is it for Higher Education?
5. How precise is it for Higher Education?
6. How complete is it for Higher Education?
7. Which issues are already covered by existing HE programs or courses?
8. Which steps would be needed to follow to introduce Trustworthy AI education in HE programs and courses?

Agenda

Subject to change, to accommodate cultural requirements:

09:00-09:30 ◦ Introductions, including presentation of the Trustworthy AI guidelines
09:30-10:00 ◦ HE program (or course) case introduction by the organisation
10:00-10.30 ◦ General Perspectives and Questions
10.30-11:00 ◦ Questions of the Assessment List
11:00-11:30 ◦ Ordering of the Requirements
11.30-12.00 ◦ Questions for Specific Areas of the Assessment List
12:00-12:15 ◦ Closing remarks

A. Introduction

Introduction of the people involved in the meeting, project - background, consent issues, description of process, and follow up steps, etc. Ensure that interview can be (voice) recorded.

B. Introduction of the Purpose of the Interview

1.   Slide deck
2.   Ensure that interview can be (voice) recorded. Make it clear that none other will have access to the recordings and that they will be deleted upon the completion of the project report.
3.   Ensure that it is understood it is not about the performance or vision of the entity but about the suitability of the trustworthy guidelines to improve Higher Education.
4.   Make it clear that the individuals will not be noted by name anywhere. Any information, e.g. their role or location will only be used in an aggregated manner.
5.   Determine whether they would like to list their AI assessment activities as part of the final report, or whether that is confidential.
6.   State the agenda for the day.

C. Education case (Education program or a course)

Discuss the higher education program (or course) with them. Allow them to present the education case. Make it clear that the scenario is meant to provide contextual information.
    Possible topics to discuss/ask:
1.   Learning outcomes.
2.   Learning outcomes vs the seven requirements of the HLEG guidelines.
3.   Teaching material.
4.   Examination methods.
5.   Heterogeneity of the students.
6.   Employability of students.

D. General Perspectives

This section is a generalised discussion of the HLEG guidelines and its assessment of AI systems during their development, deployment, and use.
1.   How would you describe the current status of "trustworthy" AI in connection with higher education in general? (e.g. national education strategies, practices in current education at your organisation.)
2.   Can you say something about the strategy of your organisation has for AI education development? (Purpose, development, administration, recent initiatives)
3.   Which of the HLEG requirements are you already teaching in your education program? Do you teach other issues related to trustworthy AI?
4.   In which education cycle should trustworthy AI education need to start?
5.   What resources should be available for trustworthy education in HE?
6.   Are there any resources that you already use for teaching aspects related to trustworthy AI?
7.   What are the minimum incentives that should be there for promoting trustworthy AI in HE?
8.   What risks and opportunities do you associate with trustworthy AI in HE?
9.   How could HE benefit from trustworthy AI?

E. Questions on the Assessment List

This section aims to ask questions on the HLEG assessment list. Try to keep the discussion within reasonable time limits.
1. In which language did you read the Guidelines and the Assessment List? (Should be asked prior to the interview, but again during)
2. In overall terms, is the assessment list useful for education purposes? Why/Why not?
3. Is it beneficial or not to make it part of an actual HE course? If so, in which form? If not, why?
4. What type of support do you need to teach the Guidelines?

F. Ordering of the Requirements

In this section, request the participants to rank the 7 Requirements (Transparency, Accountability,…) in order of their application/importance (with 1 being the highest).

Make it clear that the ordering is in terms of significance to their education given the education case and within the context of this interview. You may use the printout cards and/or remind them of the 7 requirements.

If multiple persons/roles are taking part in the interview, you may record any notable disagreements, but only one order is permitted, i.e. the organisation's position.
1. Interpretation (their own words) of each of the 7 key requirements.
2. Which TAIG requirements of the assessment list are relevant? Why/why not? In which order? Make a ranked list.
3. Why this order and why some requirements are considered less or not relevant?
4. Are there requirements in the TAIG not relevant? Why?

G. Questions for Specific Areas of the Assessment List

Take the two extremes (i.e. the highest and lowest priority) from the list produced in the previous section of the interview.
1. Which aspects are already considered in their education?
2. Why is/isn't the requirement relevant for your education?
3. Evidence of addressing it:
    a. What are the learning outcomes related to the requirement?
    b. What are the evaluation methods related to the learning outcomes?
4. Do you think this requirement is clearly outlined in the Assessment List? Could you tell me how you interpret it?
5. Which questions around this requirement are the most valuable ones for trustworthy AI education?

H. Closing remarks

A quick wrap-up of the interview.
1. Is there anything else you would like to add/ask?
2. What was most positive from this interview?

The 7 Requirements

1. Human agency and oversight
    a. Fundamental rights:
    b. Human agency:
    c. Human oversight:
2. Technical robustness and safety
    a. Resilience to attack and security:
    b. Fallback plan and general safety:
    c. Accuracy
    d. Reliability and reproducibility:

3. Privacy and data governance
    a. Respect for privacy and data Protection:
    b. Quality and integrity of data:
    c. Access to data:
4. Transparency
    a. Traceability:
    b. Explainability:
    c. Communication:
5. Diversity, non-discrimination and fairness
    a. Unfair bias avoidance:
    b. Accessibility and universal design:
    c. Stakeholder participation:
6. Societal and environmental well-being
    a. Sustainable and environmentally friendly AI:
    b. Social impact:
    c. Society and democracy:
7. Accountability
    a. Auditability:
    b. Minimising and reporting negative Impact:
    c. Documenting trade-offs:
    d. Ability to redress:

Appendix C.
Literature Review

Trustworthy **AI**

Appendix C. Literature Review

1. Methodology

The goal of the systematic review systematic review is to analyse the relevant literature in order to answer the following questions:

1. What competences and learning objectives are identified when teaching ethical aspects in HE?
2. How are these competences taught and evaluated?

We conducted the literature search on Scopus, to obtain results in a variety of disciplines. We used the following search terms, to be found in title, abstract or key:

ethics AND teaching AND "higher education" AND (competence OR competency OR skills)

We limited our search to publications from 2015 onwards and retrieved a total of 50 publications on 09/02/2021 at 15:13. We focused on individual research output, so excluded 1 book, 1 editorial, 1 extended abstract and 6 review articles (either paper reviews or curricula reviews). 4 papers were not accessible at the time of analysis. Finally, we manually excluded 8 papers whose abstract did not mention anything related to teaching skills related to ethics and 5 papers were removed upon further reading for lack of relevance (either not focused on HE or not focused on teaching aspects related to ethics). The final output **is 24 papers** which we analysed.

The papers analyse cover a wide variety of subject areas. Based on author affiliation 17 countries are covered, as well as 12 subject areas as indexed by Scopus.
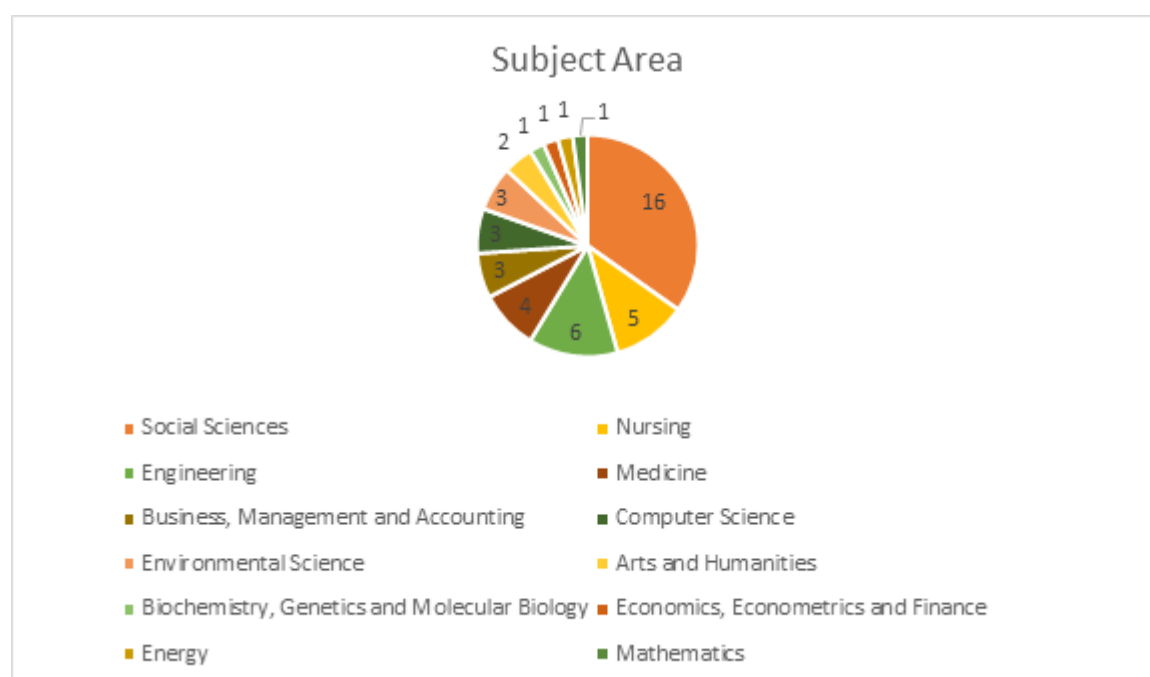
2. Overview

Publications identifying specific competences are few, although many mention that explicitly identifying competences is a pressing educational need. On the other hand, most publications propose teaching methods, with a strong focus on learning with a social component of debate and participation between students. For this reason, non-traditional teaching methods like case-studies and role-playing seem to be often proposed and studied. Much of the literature consists on exposing or evaluating how certain teaching practices were incorporated to teach ethics in specific

degrees or modules. Much of the literature emphasises the importance of incorporating different dimensions of ethics into their education. In particular, professional ethics as it refers to codes of conduct is mentioned often.

When assessing competencies related to incorporating ethical and social dimensions into Higher Education in all disciplines, results of our literature review indicate an emphasis on dual competence (Kim Brown et al., 2019; Noah & Aziz, 2020; Trobec & Starcic, 2015; Zamora-Polo & Sánchez-Martín, 2019): developing technical competence alongside the ability to understand and act according to ethical and social expectations. Although discussion on *specific* learning outcomes is notably absent, three learning goals are prevalent for demonstrating mastery of social and ethical competencies:

- Ethical appreciation/sensitivity: Identifying and understanding the ethical and moral dimensions of a situation.
- Ethical analysis: Deliberating about actions, how they relate to ethical guidelines and codes of conduct, and their possible consequences.
- Ethical decision-making/Applied ethics: Selecting and implementing a course of action in response to ethical reasoning.

Some examples of how these competences are identified can be found in the following table:

Competence	Quotes
Ethical appreciation	"When solving an ethical situation, the awareness and respect of ethical principles are not enough, the individual's role and ability to perceive the ethical dimension are also important." (Trobec & Starcic, 2015) "using technology to assess sociopolitical, ethical, and historical issues related to nursing practice" (DeSimone, 2019) "capacity to universalize principles of action, as a prerequisite for principled reasoning" (Aközer & Aközer, 2017)
Ethical analysis	"the student analyzes the pro and con viewpoints of an ethical question" (DeSimone, 2019) "reflective awareness of questions about both the "good life" and "right action" as objects of principled reasoning." (Aközer & Aközer, 2017)
Ethical decision-making	"tutors also perceived technical skills and the application skills of problem-solving and critical thinking, as important" (Dean et al., 2020) "Practice in the art of ethical decision making is the best way for ensuring that the graduates of educational leadership programs have the ethical muscle to make decisions that are moral" (Jones et al., 2020) "making and defending ethical decisions related to health-care issues" (DeSimone, 2019)

These findings squarely align with syllabi analysis, where it has been found that the most common sought outcomes for teaching Tech Ethics are variations on "recognize/critique/reason" (Fiesler, Garrett, & Beard, 2020).

The literature review reveals that ethical and moral reasoning skills are often taught through student-led methods focused on encouraging reflection and debate amongst students: case studies (Lapuzina et al., 2018), role playing (Trobec & Starcic, 2015), debate (K. Brown et al., 2019), experiential learning (Ibáñez-Carrasco et al., 2020), etc. This observation aligns with findings from other literature reviews, which emphasise the prevalence of games, role playing and case studies in Engineering and Computer Science Education (Hoffmann & Cross, 2021). Although wide-spread, there is however dissent in the literature, where some advocate for more formal training in e.g. moral philosophy, in contrast to student-led activities (Aközer & Aközer, 2017).

Across the literature there is a noticeable lack of consistent methodologies for assessing soft competences such as ethical and social awareness or understanding and application of guidelines and codes of conduct. The assessment

methods uncovered in the literature review mostly rely on self-assessment – e.g. (Ibáñez-Carrasco, Worthington, Rourke, & Hastings, 2020; Mulot-Bausière et al., 2016) – or experts' perception of student's knowledge without explicit grading criteria – e.g. (DeSimone, 2019; Lapuzina, Lisachuk, & Romanov, 2018).

3. References

Mackenzie, M. L. (2015). *Educating ethical leaders for the information society: Adopting babies from business*. *Advances in Librarianship* (Vol. 39). https://doi.org/10.1108/S0065-283020150000039010

Miñano, R., Aller, C. F., Anguera, Á., & Portillo, E. (2015). Introducing ethical, social and environmental issues in ICT engineering degrees. *Journal of Technology and Science Education*, *5*(4), 272–285. https://doi.org/10.3926/jotse.203

Trobec, I., & Starcic, A. I. (2015). Developing nursing ethical competences online versus in the traditional classroom. *Nursing Ethics*, *22*(3), 352–366. https://doi.org/10.1177/0969733014533241

Biasutti, M., De Baz, T., & Alshawa, H. (2016). Assessing the infusion of sustainability principles into university curricula. *Journal of Teacher Education for Sustainability*, *18*(2), 21–40. https://doi.org/10.1515/jtes-2016-0012

Mulot-Bausière, M., Gallé-Gaudin, C., Montaz, L., Burucoa, B., Mallet, D., & Denis-Delpierre, N. (2016). Residents' training in pain and palliative medicine: Students' assessment | Formation des internes en médecine de la douleur et médecine palliative: Bilan et suggestions des étudiants. *Medecine Palliative*, *15*(3), 143–150. https://doi.org/10.1016/j.medpal.2015.12.002

Gómez, V., & Royo, P. (2015). Ethical self-discovery and deliberation: Towards a teaching model of ethics under the competences model | Autodescubrimiento ético y deliberación: Hacia un modelo de enseñanza de la ética en el modelo por competencias. *Estudios Pedagogicos*, *41*(2), 345–358. https://doi.org/10.4067/s0718-07052015000200020

Sánchez-Martín, J., Zamora-Polo, F., Moreno-Losada, J., & Parejo-Ayuso, J. P. (2017). Innovative education tools for developing ethical skills in university science lessons. The case of the moral cross dilemma. *Ramon Llull Journal of Applied Ethics*, (8), 225–245.

Galanina, E., Dulzon, A., & Schwab, A. (2015). Forming engineers' sociocultural competence: Engineering ethics at tomsk polytechnic university. In *IOP Conference Series: Materials Science and Engineering* (Vol. 93). https://doi.org/10.1088/1757-899X/93/1/012078

Gokdas, I., & Torun, F. (2017). Examining the impact of instructional technology and material design courses on technopedagogical education competency acquisition according to different variables. *Kuram ve Uygulamada Egitim Bilimleri*, *17*(5), 1733–1758. https://doi.org/10.12738/estp.2017.5.0322

DeSimone, B. B. (2019). Curriculum Redesign to Build the Moral Courage Values of Accelerated Bachelor's Degree Nursing Students. *SAGE Open Nursing*, *5*. https://doi.org/10.1177/2377960819827086

Fernandez, F. J. I., & Martinez-Canton, A. (2019). Ethics and Transversal Citizenship in the Teaching of Science and Engineering. In *2019 9th IEEE Integrated STEM Education Conference, ISEC 2019* (pp. 104–110). https://doi.org/10.1109/ISECon.2019.8881965

Lapuzina, O., Romanov, Y., & Lisachuk, L. (2018). Professional ethics as an important part of engineer training in technical higher education institutions. *New Educational Review*, *54*(4), 110–121. https://doi.org/10.15804/tner.2018.54.4.09

Rameli, M. R. M., Bunyamin, M. A. H., Siang, T. J., Hassan, Z., Mokhtar, M., Ahmad, J., & Jambari, H. (2018). Item analysis on the effects of study visit programme in cultivating students' soft skills: A case study. *International Journal of Engineering and Technology(UAE)*, *7*(2), 117–120. https://doi.org/10.14419/ijet.v7i2.10.10968

Riedel, A., & Giese, C. (2019). Development of ethical competence in (future) nursing programs—reification and steps as a foundation for curricular development | Ethikkompetenzentwicklung in der (zukünftigen) pflegeberuflichen Qualifizierung – Konkretion und Stufung als Grundlegung für . *Ethik in Der Medizin*, *31*(1), 61–79. https://doi.org/10.1007/s00481-018-00515-0

Aközer, M., & Aközer, E. (2017). Ethics Teaching in Higher Education for Principled Reasoning: A Gateway for Reconciling Scientific Practice with Ethical Deliberation. *Science and Engineering Ethics*, *23*(3), 825–860. https://doi.org/10.1007/s11948-016-9813-y

Oliphant, T., & Brundin, M. R. (2019). Conflicting values: An exploration of the tensions between learning analytics and academic librarianship. *Library Trends*, *68*(1), 5–23. https://doi.org/10.1353/lib.2019.0028

Brown, K., Connelly, S., Lovelock, B., Mainvil, L., Mather, D., Roberts, H., … Shephard, K. (2019). Do we teach our students to share and to care? *Research in Post-Compulsory Education*, *24*(4), 462–481. https://doi.org/10.1080/13596748.2019.1654693

Jones, K. D., Ransom, H., & Chambers, C. R. (2020). Teaching Ethics in Educational Leadership Using the Values–Issues–Action (VIA) Model. *Journal of Research on Leadership Education*, *15*(2), 150–163. https://doi.org/10.1177/1942775119838297

Bates, J., Cameron, D., Checco, A., Clough, P., Hopfgartner, F., Mazumdar, S., … de la Vega de León, A. (2020). Integrating fate/critical data studies into data science curricula: Where are we going and how do we get there? In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 425–435). https://doi.org/10.1145/3351095.3372832

Dean, B. A., Perkiss, S., Simic Misic, M., & Luzia, K. (2020). Transforming accounting curricula to enhance integrative learning. *Accounting and Finance*, *60*(3), 2301–2338. https://doi.org/10.1111/acfi.12363

Zamora-Polo, F., & Sánchez-Martín, J. (2019). Teaching for a better world. Sustainability and Sustainable Development Goals in the construction of a change-maker university. *Sustainability (Switzerland)*, *11*(15). https://doi.org/10.3390/su11154224

Ibáñez-Carrasco, F., Worthington, C., Rourke, S., & Hastings, C. (2020). Universities without walls: A blended delivery approach to training the next generation of hiv researchers in canada. *International Journal of Environmental Research and Public Health*, *17*(12), 1–12. https://doi.org/10.3390/ijerph17124265

Sahin, Y. G., & Celikkan, U. (2020). Information technology asymmetry and gaps between higher education institutions and industry. *Journal of Information Technology Education: Research*, *19*, 339–365. https://doi.org/10.28945/4553

Noah, J. B., & Aziz, A. B. A. (2020). A case study on the development of soft skills among TESL graduates in a University. *Universal Journal of Educational Research*, *8*(10), 4610–4617. https://doi.org/10.13189/ujer.2020.081029