



value-**ALIGNED** socio-technical systems using large-language models (LLMs)

WP3 - Identification of societal values and user preferences in the context of the three use cases

Impact assessment of values affected

Contractual Delivery Date	28/02/2026	Actual Delivery Date	28/02/2026
Responsible Beneficiary	TUM	Contributing Beneficiary	EPFL
Security	Public	Nature	R – Document, report
Version	1	Date	28/02/2026
		Page Nb.	93

Authors

Name	Organisation	Email
Julia Li	TUM	Julia.li@tum.de
Prof. Dr. Christoph Luetge	TUM	Luetge@tum.de
Dr. Auxane Boch	TUM	Auxane.boch@tum.de

Document History

Revision	Date	Modification	Contact Person



Table of contents

Glossary.....	5
Section 1: Background and introduction	6
1. Introduction	6
2. Background	6
2.1 Human values in AI.....	6
2.2 Related work.....	7
2.3 EU societal values in AI	11
2.3.1 Values in ethical guidelines and soft-law.....	11
2.3.2 The EU AI Act.....	11
2.4 Theories about human values	12
2.4.1 Aristotelian virtue ethics	12
2.4.2 Deontological ethics	13
2.4.3 Theoretical models of human values: Schwartz and Rokeach	13
3. Translating human values into AI values	16
3.1 Project objective	16
3.1.1 Transforming human values into technical requirements	16
3.1.2 Research questions.....	17
3.2 Project framing: value-ALIGNed socio-technical systems using large-language models	17
4. Literature search	19
4.1 Grey literature and normative guidelines.....	19
4.2 Empirical literature.....	19
5. Value extraction methodology	21
5.1 Thematic analysis approach	21
5.2 Value identification	21
5.3 Value categorization.....	22
6. Results	22



6.1 Values from normative ethical guidelines.....	22
6.1.1 Preliminary value identification	22
6.1.2 Building an interactional framework of value alignment	23
6.2 Description of main themes	27
6.2.1 What kinds of qualities do we want AI to have?.....	27
6.2.2 How should AI behave towards humans?.....	28
6.2.3 How should humans behave when using AI?	28
6.2.4 What kinds of human infrastructure and processes should there be to support ethical AI use?.....	28
6.2.5 What kinds of societal goods can AI contribute to?	29
Section 2: Use-case specific values, risks and impacts	30
7. Education	30
7.1 Overview	30
7.2 Normative values in education and AI	31
7.3 Analysis of main themes from empirical literature	35
7.3.1 Academic dishonesty & fairness.....	35
7.3.2 Automated grading systems & fairness	36
7.3.3 Access to benefits of AI tools	37
7.3.4 Changes to learner and educator skills	37
7.3.5 Keeping education a neutral space.....	39
7.4 References for education related papers	41
8. Healthcare and mental health	44
8.1 Overview	44
8.2 Normative values in healthcare and AI	45
8.3 Analysis of main themes from empirical literature	52
8.3.1 Human oversight.....	52
8.3.2 Data privacy.....	55
8.3.3 Practitioner-patient relationship and “the human touch”	56



8.3.4 Transparency and trust	57
8.4 References for healthcare related papers	60
9. Journalism and online news consumption.....	61
9.1 Overview	61
9.2 Normative values in journalism and AI	62
9.3 Analysis of main themes	68
9.3.1 Trustworthiness and the perception of truth	68
9.3.2 Impacts on underrepresented groups	70
9.3.3 The changing nature of journalism	71
9.3.4 Accountability and standard setting	72
9.3.5 Impacts on local/small/regional news	74
9.3.6 Journalistic autonomy and involvement in AI development.....	75
9.4 References for journalism/online news consumption use-case	77
Section 3: Overview of findings	79
10. Bibliometric analysis of empirical literature	79
10.1 Result and discussion of bibliometric analysis results	79
10.1.1 Geographic bias towards Western Europe	79
10.1.2 Diversity of subject areas	80
10.1.3 Publications by year	81
10.1.4 Representation of women in authorship.....	82
10.1.5 Qualitative focus	83
11. Key takeaways and next steps	84
11.1 Future directions	85
11.2 Conclusions.....	86
11.3 Disclosure of AI use	87
References.....	88

Glossary

LLM – Large language model

AI – Artificial intelligence

Fiduciary relationship – A relationship with legal or ethical obligations where one party must behave in the best interests towards the other (e.g. doctor-patient relationships)

Epistemic justice – An injustice done towards somebody's capacity as a knower or knowledge-holder.

Value-alignment – Ensuring that AI systems behave according to human goals and preferences

Normative value - Aspirational standards or rules

Instrumental value – A value that is a means to another end

Deontological ethics – A school of normative ethical theory which focuses on if actions or behaviours adhere to prescribed rules

Qualitative coding – A method of qualitative analysis that labels and extracts information from raw data to make sense of relevant themes and patterns

Section 1: Background and introduction

1. Introduction

In the context of value-alignment in artificial intelligence (AI), values generally refer to an attribute, behaviour or feature of an AI system which is in line with societal aims and wellbeing (Gabriel, 2020). Similarly, human value theories have generally been framed as fundamental goals and principles that relate to human behaviour (Rokeach, 1973; Schwartz, 1992). These understandings of values are combined for the purposes of this document to integrate human values in AI research to describe principles for humans regarding the usage of AI in society and ways in which AI should “behave” according to human values within specific contexts. While these two options are not mutually exclusive, the goal of this paper is to contribute to ethical guidelines and risks and impacts which can then be applied to policy and technical requirements. Values are identified in the context of EU societal values as well as three use-cases that are part of the alignAI project; education, healthcare (mental health), and journalism (online news consumption).

2. Background

2.1 Human values in AI

Values can highlight areas of concerns that may arise and pertinent issues to consider when making decisions around AI (Jobin et al., 2019). Understanding what values mean and how best to prioritize them can serve to bolster contextual understanding about strategic direction and potential threats to societal wellbeing.

Broad, overarching human values are widely used to describe norms for AI systems in guidelines, regulations and policies (Corrêa et al., 2023). These abstract values can provide information about the ethical orientation of institutions on important policy items such as best practices, requirements and future directions (Jobin et al., 2019). A unified understanding of

what values to prioritize can be used as a reference for AI practitioners, private companies, public-sector organizations, and institutions on how to act in solidarity with an agreed upon vision for ‘ethical AI’ (Jobin et al., 2019).

2.2 Related work

Recent research has shown that human values of AI in available guidelines including those found in private company reports, regulation(s) and policy documents have begun to show a degree of convergence on several ethical principles (Corrêa et al., 2023; Fjeld et al., 2020; Hagendorff, 2020; Jobin et al., 2019). While this itself does not indicate categorical global agreement on what ‘ethical AI’ might look like, it does signal that there may be common opportunities, normative standards and risks of AI (Jobin et al., 2019).

Four key research papers utilizing textual analysis have extracted several overlapping human values and principles from AI ethics related documents worldwide. While other studies may be available, this list does not intend to be exhaustive in terms of research on guidelines and human values.

In a scoping review of 84 ethics guidelines coming largely from the USA, UK, Japan, Germany, France and Finland, Jobin et al. (2019) found eleven overarching principles.

- *Transparency*
- *Justice and Fairness*
- *Non-maleficence*
- *Responsibility*
- *Privacy*
- *Beneficence*
- *Freedom and autonomy*

- *Trust*
- *Sustainability*
- *Dignity*
- *Solidarity.*

Hagendorff (2020) found that *Accountability, Privacy* and *Fairness* appeared in 80% of 22 guidelines surveyed in a wide variety of private and public organizations between 2015-2020, with *Accountability, Explainability, Privacy, Justice, Robustness* and *Safety* appearing often in the context of implementable technical solutions (2020).

In a review of 38 organizational and institutional documents, Fjeld et al. found eight key themes which also overlap with other researchers' findings (2020).

- *Privacy*
- *Accountability*
- *Safety and Security*
- *Transparency and Explainability*
- *Fairness and Non-discrimination*
- *Human Control of Technology*
- *Professional Responsibility*
- *Promotion of Human Values*

Lastly, Corrêa et al. located 17 principles in a review of 200 documents worldwide in which four values were noted as being most prominent (2023).

- *Transparency/Explainability/Auditability*
- *Reliability/Safety/Security/Trustworthiness*
- *Justice/Equity/Fairness/non-discrimination,*

- *Privacy, Accountability/Liability*

The presence of overlapping themes found by researchers looking at AI ethics documents across the globe suggests that there is a growing shared understanding of what high-level values are at play in AI ethics and human-value alignment (Figure 1).

Figure 1

Table containing spectrum of values found in Jobin et al. (2019), Hagendorff (2020), Fjeld et al. (2020), Corrêa et al. (2023)

Ethical Theme	Jobin et al. (2019)	Hagendorff (2020)	Fjeld et al. (2020)	Corrêa et al. (2023)
Transparency	<input checked="" type="checkbox"/>			
Transparency / Explainability / Auditability				<input checked="" type="checkbox"/>
Transparency/ Explainability			<input checked="" type="checkbox"/>	
Explainability		<input checked="" type="checkbox"/>		
Privacy	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Privacy / Data Protection				
Accountability / Liability				<input checked="" type="checkbox"/>
Accountability		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	
Responsibility/Accountability	<input checked="" type="checkbox"/>			
Professional Responsibility			<input checked="" type="checkbox"/>	
Justice		<input checked="" type="checkbox"/>		
Justice / Fairness / Equity	<input checked="" type="checkbox"/>			
Justice / Fairness / Equity / Non-discrimination				<input checked="" type="checkbox"/>
Diversity/inclusion/pluralism/accessibility				<input checked="" type="checkbox"/>
Fairness / Non-discrimination			<input checked="" type="checkbox"/>	
Fairness		<input checked="" type="checkbox"/>		
Robustness		<input checked="" type="checkbox"/>		
Safety		<input checked="" type="checkbox"/>		
Safety / Security			<input checked="" type="checkbox"/>	

Trust / Trustworthiness / Reliability / Safety / Security				
Reliability / Safety / Security / Trustworthiness				<input checked="" type="checkbox"/>
Trust	<input checked="" type="checkbox"/>			
Truthfulness				<input checked="" type="checkbox"/>
Non-maleficence (Do no harm)	<input checked="" type="checkbox"/>			
Beneficence/ Non-maleficence				<input checked="" type="checkbox"/>
Beneficence (Promote Wellbeing)	<input checked="" type="checkbox"/>			
Promotion of Human Values			<input checked="" type="checkbox"/>	
Human-centeredness/alignment				<input checked="" type="checkbox"/>
Human Control of Technology			<input checked="" type="checkbox"/>	
Freedom / Autonomy	<input checked="" type="checkbox"/>			
Freedom/autonomy/democratic values/technological sovereignty				<input checked="" type="checkbox"/>
Sustainability (Environmental)	<input checked="" type="checkbox"/>			
Sustainability (Social/Intergenerational/Environmental)				<input checked="" type="checkbox"/>
Dignity	<input checked="" type="checkbox"/>			
Dignity / Human rights				<input checked="" type="checkbox"/>
Children's and Adolescent Rights				<input checked="" type="checkbox"/>
Solidarity (Labor rights)	<input checked="" type="checkbox"/>			
Labor rights				<input checked="" type="checkbox"/>
Human formation/education				<input checked="" type="checkbox"/>
Intellectual property				<input checked="" type="checkbox"/>
Cooperation/fair competition/open source				<input checked="" type="checkbox"/>

Note. Non-exhaustive list of main values extracted listed *ad verbatim* as they appear in the original articles

2.3 EU societal values in AI

While a growing body of literature has looked at high-level human values in AI ethics across the globe, the specific nuances of value-alignment in certain regions such as in the EU has yet to be extensively investigated (Hagendorff, 2020). In research studies mapping human values and principles of AI, the EU has often emerged as one of the most prolific regions in terms of producing AI ethics guidelines and documents (Corrêa et al., 2023; Felkner et al., 2024; Fjeld et al., 2020; Hagendorff, 2020; Jobin et al., 2019).

2.3.1 Values in ethical guidelines and soft-law

There are a multitude of normative sources describing human values and AI in the EU. Ethical guidelines and soft-law documents have applied a great deal of normative force in the EU. For instance, “AI4People’s Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations” which focuses on European institutions and future AI risks was fully adopted by the High-Level Expert Group on AI (Floridi et al., 2018). Combining and synthesizing 6 other sets of 47 existing principles from other reputable, multi-stakeholder organizations, the identified standards and principles were contextualized for the EU and form the basis of recommendations for the ethical design, development and implementation of AI (2018). The authors, which comprise of scientific experts, emphasize that they drew from the four bioethical principles of *beneficence*, *non-maleficence*, *autonomy* and *justice*, adding *explicability* (incorporates *intelligibility* and *accountability*) to form a unified framework (2018). These principles encompass the opportunities and risks associated with AI technologies and their contribution to what they coin “Good AI Society.”

2.3.2 The EU AI Act

The EU’s approach to AI governance also reinforces human values with the EU AI Act ((EU) 2024/1689 regulation. The Act operates on a risk-based tiering system that enforces compliance with legal and financial penalties, drawing its legal force from fundamental human rights (Pham

& Davies, 2025). Human values that are defined in the regulation and through special requirements for high-risk AI systems developed considering the Act are relevant ethical reference points which are central to the EU “Trustworthy AI” approach (Billah et al., 2025).

- *Data governance*
- *Record-keeping*
- *Transparency*
- *Human-oversight*
- *Accuracy*
- *Robustness*
- *Cybersecurity*

These requirements are connected to other human values and requirements found in soft-law and policy documents (Balcioglu et al., 2025). In accordance with the risk-based nature of the EU AI Act, the requirements set out in Trustworthy AI are distinctly directed towards informing and regulating the development of AI in the EU context.

2.4 Theories about human values

2.4.1 Aristotelian virtue ethics

The goal to understand and define human values have roots in foundational theories of human values and ethics. Early discussions of human values can be found in Aristotle’s virtue ethics (Kraut, 2022), who identified virtues as character traits enabling individuals to achieve “eudaimonia” or flourishing, a state of happiness or living well (Kraut, 2022). These ethical values featured in Aristotle’s writings, such as “justice” and “friendship” are not entirely congruous with the framing of values found in currently available AI ethics guidelines (Dobre & Dobre, 2021; Kraut, 2022). However, they paint a familiar picture of traits, experiences and societal goods that are desirable to achieve in life (Dobre & Dobre, 2021; Kraut, 2022).

2.4.2 Deontological ethics

Deontological ethics pioneered by philosophers such as Kant describe a system of moral principles and maxims by which to judge the moral quality of actions which align or do not align with them (Kant, 2018/1785). The framework offered by deontology can inform a way of conceptualizing modern ethical guidelines that allude to rights and duties such as in the case of human rights or bioethical duties of physicians towards patients (D'Alessandro, 2025; Kant, 2018/1785). Furthermore, central maxims of Kant's philosophy such as treating humans as individuals with autonomy or "treating humans as ends in themselves and not merely as a means" has the effect of embodying certain values such as *human dignity* and *human autonomy* which also appear in AI ethics (D'Alessandro, 2025; Kant, 2018/1785).

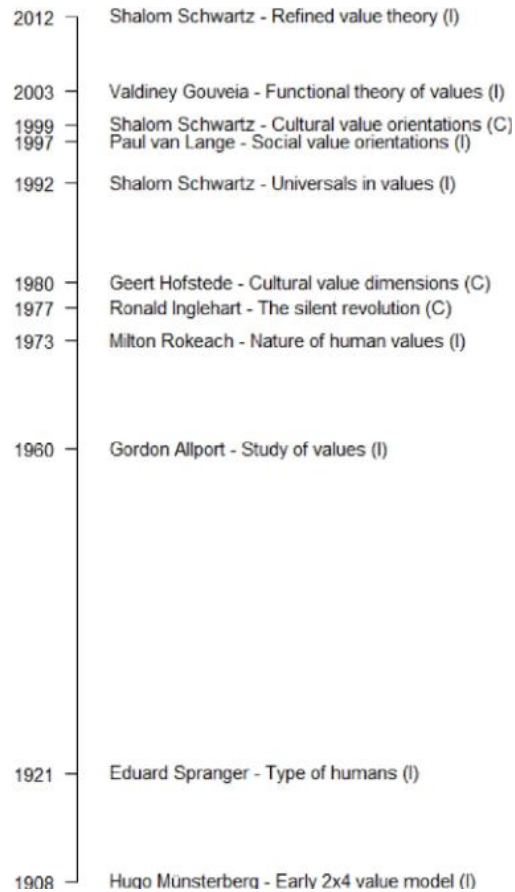
2.4.3 Theoretical models of human values: Schwartz and Rokeach

Building upon previous value theories, theorists in the 20th and 21st century in the field of psychology attempted to establish systems of basic human values and beliefs which could characterize the human experience (Figure 2) (Hanel et al., 2018).

For instance, Rokeach's value theory asserted that there may be an observable connection between behaviour and stated values (1973). Rokeach's experiments in social psychology on 366 Michigan State University students showed evidence that certain values like *equality* and *freedom* could be enduring and determine behaviours throughout the course of up to two years (Rokeach, 1971). His proposition of the existence of terminal values, which are abstract and guiding principles, and instrumental values, thought of as ways to reach those goals (Rokeach, 1973), were eventually built upon by Schwartz (1992) and other theorists examining the motivational quality of human values.

Figure 2

Historical overview of selected important contributions to human value research.



Note. Reprinted from Hanel, P. H. P., Litzellachner, L. F., & Maio, G. R. (2018). An Empirical Comparison of Human Value Models. *Frontiers in Psychology, 9*.

<https://doi.org/10.3389/fpsyg.2018.01643>

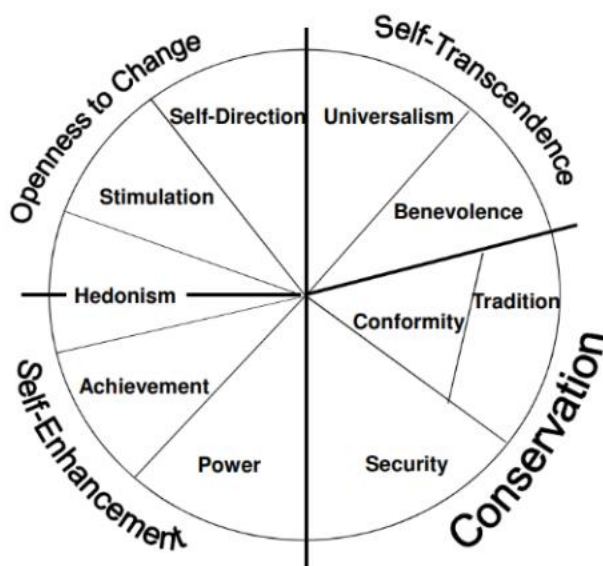
Schwartz's contribution is most recognizable as the circumplex value system describing key motivating life principles that are supported by underlying motivations (Schwartz, 1992) (Figure 3). Four higher-order values including *Openness to Change*, *Self-transcendence*, *Self-enhancement* and *Conservation* were theorized to characterize broad beliefs which would then correspond or conflict with lower-order values like *Hedonism* or *Benevolence*. He posited 19

value types based on analyses of self-reported values from over 20 countries that formed the foundation of his theory of universal human values (Roccas et al., 2002; Schwartz, 1992, 2012) (Figure 3). Other theories have gone on to elaborate on Schwartz’s circular value framework, such as Gouveia’s theory of functional values which further separate Schwartz’s values into motivational and basic values (Gouveia et al., 2014).

Amongst others, these theories form the foundation of value theory and describe a motivational value structure that also corresponds to the way that values in value-alignment are often conceptualized.

Figure 3

Schwartz’s “Theoretical model of relations among ten motivational types of value”



Note. Reprinted from Schwartz, S. H. (2012). An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture*, 2(1). <https://doi.org/10.9707/2307-0919.1116>

3. Translating human values into AI values

3.1 Project objective

However, all theories of fundamental human values generally apply to motivations and goals as embodied and experienced by people and society. Human value alignment in AI is typically conceptualized as aligning the behaviour, development and/or implementation of AI with desirable traits, end-states, or other goods (Gabriel, 2020). Although some human values such as Schwartz's *benevolence* and Rokeach's *freedom* and Maslow's *safety* also re-emerge as values discussed in contemporary AI ethics, they need to be transformed into definitions that are applicable for AI governance and technical value alignment.

3.1.1 Transforming human values into technical requirements

Technical value-alignment methods require methods to define and quantify desirable values to embed into AI systems like large language models (LLMs), and how to navigate conflicts between values. There has been some recent work on frameworks to outline ways to fine-tune LLMs according to societal values, such as COUPLE (Guo et al., 2025) and ValueCompass (Shen et al., 2025) as well as datasets like ValueEval24 (Yeste & Rosso, 2026) that can assist in benchmarking the performance of LLMs against human values and preferences. The body of literature on societal values on LLMs and AI, from empirical, theoretical, and normative perspectives is also growing to accommodate new methods of testing whether LLMs align with desired values and how to mold LLMs to adhere to values and principles important to humans.

Despite the growing body of relevant literature providing context on both human values and values in AI ethics, there is an overall lack of research on methodological and philosophical bridging which can connect human value theories and core values that have been identified regarding what is useful for a large language model (LLM) to embody (Han et al., 2022).

Connecting EU-specific values such as those in Trustworthy AI, and values found from empirical

research conducted in EU populations and their perspectives on AI adds an additional layer of complexity.

This document draws from a foundation of value frameworks developed by Schwartz and Rokeach to provide an overview of values and their sub-values, conflicts/trade-offs, opportunities, and risks and impacts in EU societal values on LLMs and AI. By using a simple, reproducible framework for value identification and mapping values from a diverse range of normative and empirical literature, this research provides a list of values and definitions that can be used to develop technical and governance frameworks in the future.

3.1.2 Research questions

Main RQ: What are EU societal values on LLMs and AI systems in normative and empirical literature and how are they defined?

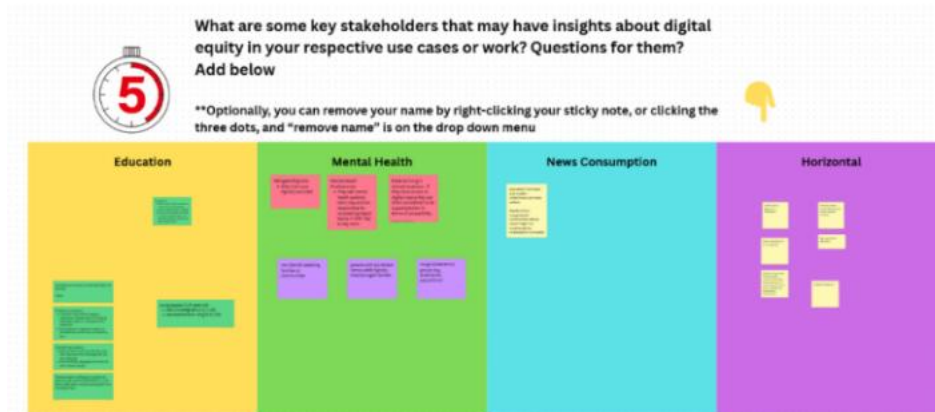
Sub RQ1: What are the relevant risks and impacts of AI and LLMs on the general population and vulnerable groups?

3.2 Project framing: value-ALIGNED socio-technical systems using large-language models

The identification of values is further subdivided into three use-cases that are part of the Horizon Europe, Marie Skłodowska-Curie Actions (MSCA)-funded value-ALIGNED socio-technical systems using large-language models (LLMs) (alignAI) project. This document serves as an initial risk and impact assessment for values in three LLM and AI use-cases in education, mental health and online news consumption. Healthcare is used as a proxy for mental health and journalism is used as a proxy for online news consumption to gain relevant insights from the preliminary data gathered and analyzed.

Figure 4

Screenshot of exercise conducted during the first doctoral seminar in conjunction with 16 other DCs to identify stakeholders and main considerations about digital equity



In the 10 months leading up to the deliverable, the author and main beneficiary of this deliverable (DC1 at the Technische Universitaet Muenchen) engaged in interdisciplinary discussions with colleagues to gain insight into relevant considerations in each use-case as well as how values can be translated into technical requirements, informing the priorities listed below (Figure 4). Key priorities about value identification that were expressed through these collaborative activities included:

- 1) Relevant values per use case
- 2) Relevant conflicts and trade-offs
- 3) Translating values into technical requirements
- 4) Defining values and
- 5) Identifying possible impacts on vulnerable and marginalized groups.

Future steps and collaborations include:

- a) Cross-referencing values identified in each use case with relevant stakeholder interviews
- b) Identifying how the EU AI Act protects impacted values and vulnerable groups

- c) Translating values and preferences into relevant technical requirements
 - a. E.g. codifying and quantifying how to navigate value conflicts, using identified values as “ground truth” for exploring fine-tuning methods

4. Literature search

4.1 Grey literature and normative guidelines

Two groups of literature were identified. The first group of articles was a collection of grey literature (n=13) including soft policy, ethical guidelines, reports and human value theory collected purposively according to their relevance to EU AI ethics and normative ethical guidelines for the three use-cases.

Trustworthy AI (High-Level Expert Group on Artificial Intelligence, 2019) and AI4people’s Ethical Framework (Floridi et al., 2018) were identified as key normative ethical frameworks with overarching values that would be relevant to all of the use-cases. Then, each use-case was matched with corresponding ethical guideline documents on AI and basic normative ethical guidelines for the professions.

4.2 Empirical literature

Empirical literature was taken from a preliminary sample of peer-reviewed studies published between 2017-2025 that were identified as part of a systematic review on EU societal values.

Overall, 21 articles pertaining to three use-cases in education, healthcare (mental health) and journalism (online news consumption) were selected from a sample of 50 full-texts from the systematic review (Figure 5). Five more relevant articles focusing on journalism were identified in a Google scholar search conducted in February 2025. All empirical literature assessed in this analysis falls within the eligibility criteria established in the systematic review (Figure 6).

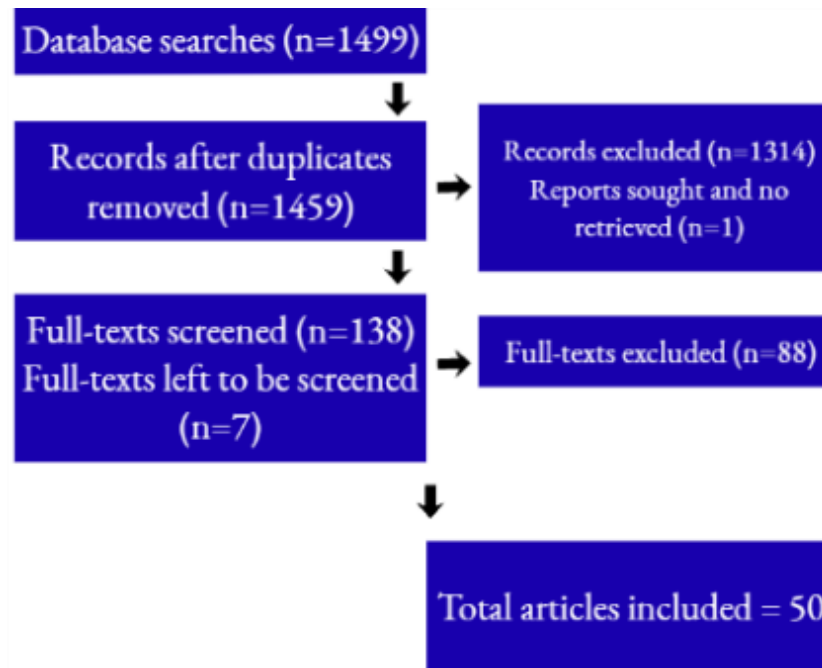
Figure 5

Inclusion criteria

Sample	Populations and individuals located in European Union member states
Phenomenon of Interest	European societal values
Design	Published literature of empirical research designs
Evaluation	Types of values, ethics, ethical norms, perspectives, attitudes, opinions, and/or standards
Research type	Literature containing qualitative, mixed-methods, quantitative results published from beginning of 2017- 2025

Figure 6

PRISMA diagram of preliminary results from systematic review



Note. 21 articles were included from the sample of 50 articles due to their relevance to education, healthcare (mental health), and journalism (online news consumption)

5. Value extraction methodology

5.1 Thematic analysis approach

Drawing from Johanna Briggs Institute (JBI) guidelines on qualitative evidence synthesis methods (Aromataris et al., 2024; Lizarondo et al., 2020; Lockwood et al., 2015), insights were collected from documents using a JBI methodological approach to Textual Evidence Systematic Reviews. First, the documents were qualitatively coded along with corresponding excerpts, the source document, related values, and other relevant information such as document type, number of participants, location of study, and methodology details. Codes were iteratively grouped into categories throughout the process and continuously refined to reflect emerging categories. After collecting themes from the relevant articles, the textual evidence was arranged in terms of the claims, the data (grounds) used to support the claim, and the warrant or connection between the data and claim (Aromataris et al., 2024). Then, the values were collected and further refined into subcategories and overarching themes.

5.2 Value identification

Drawing from hierarchical value frameworks (Rokeach, 1973; Schwartz, 2012), values were qualitatively extracted from relevant policy and academic documents in a semi-deductive manner. A flexible codebook of keywords from academic literature on human values in AI guidelines (Corrêa et al., 2023; Fjeld et al., 2020; Hagendorff, 2020; Jobin et al., 2019) were used to label excerpts from the texts, which were then revised, merged and re-labelled throughout the course of the coding. Predefined categories which separated high-level values from secondary/instrumental values and descriptions which serve to further contextualize the values were used as a template for coding.

5.3 Value categorization

The concept of end-state values and requirements or instrumental values draws from theories mentioned in 2.4.3 such as Rokeach's value inventory which describes a similar system of instrumental and terminal human values (1973). In separating values between those that are normative and those that are more descriptive, the distinction can be made between sub-goals that are descriptive of actions or behaviours and more abstract end-states or terminal values. The resulting inventory of sub-goals can be used as a framework for defining relevant values and their dimensions.

During the extraction, values were categorized according to complexity and function. Main values are defined as those that are abstract concepts, end-states, or goals. Sub-values are defined as actions, attitudes, principles, instrumental goals or preferences that provide more information about the core value in a particular context. This methodology was applied across all the documents to identify and collect values and what they may mean in specific contexts.

6. Results

6.1 Values from normative ethical guidelines

6.1.1 Preliminary value identification

Firstly, Trustworthy AI (High-Level Expert Group on Artificial Intelligence, 2019) and AI4People (Floridi et al., 2018) guidelines were broadly coded for main themes. Values identified in the Trustworthy AI and AI4People documents (Figure 12) closely align with those previously found in the literature, overlapping considerably with each other on major concepts such as *security*, *transparency*, *freedom*, and *non-discrimination* from reviews on AI ethics guidelines found globally (Corrêa et al., 2023; Fjeld et al., 2020; Hagendorff, 2020; Jobin et al., 2019) (Figure 1). Although some values are more broadly generalizable, others are highly specific to areas like *regulation*, *technical robustness*, or *reporting processes*.

containing subthemes. After reorganization, categories of values were separated based on what or who they are describing, and whether the value is a requirement or an abstract end-state

The codes were arranged under sub-questions presented by Shen et al.’s concept of a Bi-directional Human-AI alignment Framework (2024) which is also drawn from adapted Schwartz Theory of Basic Values. Shen et al (2024) state that the focus on designing AI systems to facilitate human understanding overlooks the goal of fostering AI literacy to support AI engagement and collaboration. Instead of outlining solely values and sub values, this framework focuses on the interactional nature of human values to outline technical and societal techniques for AI value alignment (Shen et al., 2024).

Expanding this framework, the values were further coded to characterize the following (Figure 8):

- Ideal behaviours AI should have towards users (instrumental)
- How humans should behave in relation to the use of AI (instrumental)
- What kinds of end-state values we want AI to embody (terminal)
- What societal goods can AI help humans achieve (terminal), and
- What kinds of societal processes promote these (instrumental)

Figure 8

Framework for societal value alignment for AI from values drawn from Trustworthy AI and AI4People principles

How should AI behave in relation to humans?	
Reproducible	Behaves in a way which is reproducible across contexts.
Avoids unfair bias	Avoids harmful and unfair biases, especially those that perpetuate harmful societal stereotypes and prejudices
Trustworthy	Able to be trusted to act according to human values (keep data private, be benevolent, not create harm, etc...)

Socially aware	Can behave reliably in line with societal and human interests, avoiding offensiveness, non-ethical behaviours, and other anti-social behaviours
Aligned judgment	Can make the correct assessment and decisions according to human preferences and values.
Ethical	Behaves according to ethical norms and principles.
Benevolent	Promotes human and environmental well-being.
Comprehensible	Can be understood by the relevant stakeholder(s)/user(s).
Accurate	Makes decisions that are close to what is true or correct.
Respect for human autonomy	Does not undermine human self-determination and upholds the autonomy of users and stakeholders.
Non-malevolence	Preventing and avoiding actions which cause harm to humans and/or the environment.
Procedurally fair	Ensuring that outcomes are reached in a fair, just and explicable manner, and that there are mechanisms and pathways for accountability and redress for decisions made by AI systems.
Respect for informed consent	The user needs to be made aware that they are interacting with an AI system and do so with the understanding of its abilities and limitations
Privacy	Ensures the privacy of the user through technical safeguards, specifications, and behavioural characteristics.
How should humans behave in relation to using AI?	
Lawful	Respect laws and regulations around AI system development, engagement, and distribution.
Ethical[ly]	Respect ethical norms and principles when engaging with AI systems, avoid malicious behaviour and consider the ethical impacts of one's actions.
Proportionality	Humans should weigh the costs and benefits of developing, deploying and using AI systems, especially regarding trade-offs and competing interests.
What kind of societal processes should we have to promote these goods?	
Data governance	Having a system for data management throughout the lifecycle of the AI system.
Stakeholder participation	Involving and engaging stakeholders meaningfully throughout the lifecycle of the AI system.
Accountability	Mechanisms that allow the responsible part(ies) to be accountable for the AI system, its outcomes and other impacts
Auditability	Technical and governance mechanisms that allow the AI system's operations to be evaluated
Redress	Mechanisms that ensure that individuals and groups, particularly marginalized individuals and groups, have pathways to seek remedy or compensation
Reporting	Processes are set in place that allow the AI system and its potential and actual impacts to be identified, assessed and documented.

Risks and impacts assessments	Ensuring that all necessary and proportional steps are taken (risk and impact assessments, testing, evaluation, reporting, auditing, etc..) to minimize the potential of harmful impacts
Diverse and inclusive development	Teams of AI practitioners should be inclusive and diverse, reflecting different backgrounds and skillsets.
What kind of qualities do we want the AI to have?	
Resilient	Withstands and recovers from challenges.
Secure	Protects people or the environment from intended harms (e.g. attacks, malicious activity).
Safe	Protects people or society from unintended harms.
Reliable	Maintains consistency in performance and behaviour over a variable of environments such that it performs as it is expected to.
Robust (technically)	The system is technically robust and ascribes to technical specifications that makes it trustworthy.
Auditability	The system allows for auditing and accountability.
Allows for human intervention	The system has features that allow for human intervention and agency.
Explainable	The logic of decisions made by the AI system should be made clear and understandable to the user and stakeholders.
Traceable	The behaviour of the AI system is traceable such that one can determine relevant actors, actions, and other measures important for transparency.
Substantively fair	The system engages in "fair" decision making processes ensuring that benefits and costs are distributed equally and justly amongst stakeholders and users.
Accessible	Accommodating for differences in individuals and accessible regardless of disability status or personal characteristics.
Environmentally sustainable	The system should limit its negative impact on the environment through minimizing its power and resource consumption
Transparency	The AI system's data, system specifications and AI business models should be transparent to stakeholders and users.
Explicable	AI systems should behave in a way that allow humans to understand and hold to account decision-making processes
What societal goods can AI help humans achieve?	
Privacy	Having freedom from intrusion and control over one's personal information.
Diversity	Fosters diversity by being accessible to all and actively engaging relevant stakeholders throughout the lifecycle.
Planetary wellbeing	Contributes to holistic planetary wellbeing, contributing to biodiversity, nature, and environmental sustainability

Democracy	Contributes to democratic processes and fair political processes, minimizing harm from disinformation, misinformation and political interference.
Socially sustainable	Takes into consideration the wellbeing of human society including future generations.
Equity	Prioritizing the rights and wellbeing of vulnerable and marginalized populations in all aspects of AI system development, deployment and use.
Digital inclusion	Bridging the gap in access to digital services which affect the quality of life and wellbeing of individuals and populations.
Freedom	AI systems should foster freedom (of speech, of thought, of association etc..), along with promoting other fundamental rights.
Human dignity	AI systems should promote human dignity and fundamental human rights, centering human goals and desires in its functions.
Justice	The development of AI systems should ensure the just distribution of risks and benefits in society.
Beneficence	The development of AI systems should further the wellbeing of individuals, the planet, and public good.

Note. Values and descriptions derived from combined codes from TrustworthyAI and AI4People.

6.2 Description of main themes

6.2.1 What kinds of qualities do we want AI to have?

These values are broad and denote a range of characteristics, abilities and capacities of an AI system. They are described as nouns which refer to broad concepts that are linked to various system capabilities. As well, many of these values describe technical features that are not oftentimes values used in human value theory, including *resilience, security, safety, reliability, robustness (technical), traceability, transparency, auditability, reproducibility, comprehensibility, accuracy, privacy, explainability, and explicability*. *Accessibility* and *fairness* are closely connected, referring to the idea that systems themselves should be constructed in a way that benefit all, especially vulnerable groups. Most of these values are directed towards the effects that systems have on humans except for *environmental sustainability* which refers to a broader set of impacts on humans, non-humans and the natural environment.

D3.1 Impact Assessment of Values Affected

6.2.2 How should AI behave towards humans?

The second categories of values describes the ideal behaviour that AI systems would have towards humans. These values define the quality of social interactions rather than the technical abilities of AI but have a degree of overlap as well. They are also values that may also be assigned to humans and describe desirable social interactions. For instance, *avoiding unfair bias, trustworthiness, social awareness, good judgment, ethicality, benevolence, respecting human autonomy, non-malevolence, fairness, respect for informed consent*. Notably, *allowing for human intervention* is a feature of AI that is distinctly different from a human-human interaction and is closely linked to the idea of *respecting human autonomy*.

6.2.3 How should humans behave when using AI?

On the flipside, humans should also embody values when using AI in order to promote personal and societal wellbeing. These values are more closely associated with interpersonal wellbeing and making decisions that respect societal norms and ethics. *Lawfulness*, and respecting laws and regulations around AI systems, *ethicality*, respecting ethical norms and principles when using AI, and *proportionality*, weighing the costs and benefits of using AI such as values like environmental sustainability vs. utility, and transparency vs. privacy.

6.2.4 What kinds of human infrastructure and processes should there be to support ethical AI use?

These values refer to specific types of societal structures or processes that are critical for supporting ethical behaviour when using AI as well as promoting societal goods. Further, these types of values are largely legalistic and administrative in nature, referring to processes and procedures embedded within human governance that help standardize AI ethics and ensure societal wellbeing. *Data governance* is a major value that encompasses a host of systems which ensure that data is private, safe, and used responsibly throughout the lifecycle of an AI system. Similarly, *accountability, auditability, reporting, risks and impacts assessments and*

redress refer to measures to ensure that AI practitioners and users are adhering to standards for responsible AI use and dissemination. *Diverse and inclusive development* or ensuring that teams of AI practitioners reflect different backgrounds and skillsets, and *stakeholder participation*, point to measures to make sure that AI development reflects the needs of diverse populations and groups within society.

6.2.5 What kinds of societal goods can AI contribute to?

Drawing from fundamental human rights as well as Rokeach's idea of terminal values that humans strive to achieve (Rokeach, 1973), this category of values refers to abstract end-states that can contribute to the wellbeing of human society broadly. For instance, *diversity, democracy, equity, freedom, human dignity, justice, beneficence, and planetary wellbeing* can be thought of goods within themselves. There are values related to rights such as the right to *privacy* and the right to *equity and equality* which are enshrined in the Universal Declaration of Human Rights (UDHR).

Section 2: Use-case specific values, risks and impacts

Drawing on literature relevant to education, healthcare, and journalism, values were identified, defined and analyzed according to their related risks, impacts and opportunities. This section contains use-case specific information from both grey literatures containing normative values and ethical guidelines in each respective sector as well as recent empirical studies.

7. Education

A total of 13/50 articles focused on education as a topic.

10/13 included students, 2/13 included students and professionals, 1/13 included students and educators, and 3/13 included only educators or professionals. Altogether, the studies included 723 participants, with studies ranging from 5-234 participants. 3/13 were quantitative, 1/13 was mixed methods, and 9 were qualitative.

14 countries were represented: Serbia (n=2), Romania (n=1), Portugal (n=1), Spain (n=2), Poland (n=1), Sweden (n=2), Austria (n=2), Finland (n=2), Estonia (n=1), Bulgaria (n=1), Denmark (n=1), Germany (n=2) and Netherlands (n=2) with 6/13 articles with participants from more than one country.

7.1 Overview

Amongst other challenges and opportunities, LLMs and AI systems introduce substantial shifts to the way students learn, and the skills needed for the future. While education has historically been associated with knowledge-gathering, the potential that AI systems have to automate many types of cognitive labour is putting the spotlight on other skills such as creativity and critical thinking which will enable humans to oversee and use AI systems (UNESCO, 2019).

It is for this reason that the impacts of AI on the labour market are a key consideration for educators and stakeholders who are preparing learners for potential shifts in the labour market and education in the coming years (Pavlova & Slavov, 2025). At the same time, there is uncertainty about how and if these changes will manifest, and what skills are necessary for students to learn and retain (Prather et al., 2023). Rapidly expanding possibilities for AI use cases in education are also changing expectations for the skillsets that educators require (Figueras et al., 2024).

More than ever, gaining basic digital literacy skills and knowledge about AI is crucial for both educators and learners to adapt to new technologies and ways of accessing knowledge (Bucea-Manea-Țoniș et al., 2022; Cernadas & Fernández-Delgado, 2021; Floridi et al., 2018; Prather et al., 2023; UNESCO, 2019).

7.2 Normative values in education and AI

Ethical frameworks such as the Beijing Consensus on Artificial Intelligence and Education (UNESCO, 2019) and AI4People draw from traditional pedagogical values that emphasize the duty of the teacher to the student/learner. However, there are important considerations related to AI that are reshaping the landscape of education. Defining best practices for the use of AI in the classroom, the skills at risk due to AI use and how to best prepare students for a changing labour market are critical pillars for responsible AI use (Figure 9).

Figure 9

Normative values in education and AI based on a thematic extraction of grey literature

Function	Values
What describes the ideal behaviour of an AI system?	Benevolent ² Linguistically inclusive ¹ Non-discriminatory ¹ Gender inclusive ¹ Non-malevolence ¹

What kind of qualities do we want AI to have?	Secure ¹ Lawful ¹ Ethical ¹ Robust (technically) ¹ Equitable ¹ Transparent ¹ Accessible ¹
What are the current policy priorities in the scope of AI development in education	Systemic collaboration ¹ Administrative digitization ¹ Revitalize teaching models ¹ Innovation ¹ Evidence-based technological change ¹ Remote assessment ¹ Develop AI talent ¹ Personalized learning ¹ Researching ethical issues in AI ¹ Globally equitable AI development ¹ Uphold human rights ¹ Close digital gaps ¹
What kind of processes should we have?	Financial management ¹ Auditability ¹ Testing and evaluation ¹ Ethical development of AI ¹ Evidence-based policy making ¹ Learning equity ¹
What kind of things should we be striving towards collectively (societally)	Humanism ³ Privacy ² Digital equity ¹ Open science ¹ Respect for human autonomy ¹ Tolerance ³ Societal equality/equity ³
What kinds of values do we want to instill in children in the scope of AI use in education?	Creativity ² Learner agency ² AI literacy ¹ Gender equity ¹ Achievement ¹ Adaptability to changing labour conditions ¹ Digital literacy ¹

Note. Based on a thematic extraction of

(1) UNESCO. (2019). *Beijing Consensus on artificial intelligence and education*. United Nations Educational, Scientific and Cultural Organization. †
<https://unesdoc.unesco.org/ark:/48223/pf0000368303>

(2) Miao, F., Holmes, W., Ronghuai, H., & Zhang, H. (2021). AI and education: Guidance for policy-makers. <https://unesdoc.unesco.org/ark:/48223/pf0000376709>

(3) OECD. (2021). *State of implementation of the OECD AI Principles: Insights from national AI policies* (No. 311; OECD Digital Economy Papers). Organisation for Economic Co-Operation and Development (OECD). <https://doi.org/10.1787/1cd40c44-en>

Normative values extracted from the selection of 13 empirical studies contain similar themes with a larger majority of values directed towards desirable skills and processes that allow AI to be integrated responsibly into the classroom (Figure 16). Values such as *Digital literacy* and *Human oversight* which can also be seen in other use-cases such as Healthcare, are contextualized for the classroom setting with sub-values that describe ways of achieving those goals. It is possible that these values differ from those extracted from the grey literature due to the fact that empirical studies provide the opportunity to garner stakeholder perspectives from educators and learners who are familiar with day-to-day activities in the job. Therefore, more abstract concepts such as *Benevolence* and *Open Science* could be less immediately important for the participants who generated the insights included in the empirical literature.

Figure 10

Normative values extracted from a selection of empirical studies

Normative values extracted from empirical articles	Sub-values
Digital Literacy	Computing skills
	Basic coding skills
	Understanding of digital technology
Regulations and Standards	Legal regulations Tech regulations Ethical norms Awareness of legal ramifications of AI use
Trust	Neutral funding bodies Content credibility Reduce bias from developers and stakeholders AI validation AI documentation AI testing Verifiability

Human oversight	<ul style="list-style-type: none"> Multistakeholder responsibility Adaptive systems Ethical alignment Human intervention Safety Human control Adaptation to human not adaptation to machine Human in the loop Clear guidance Respect for subject-specific professional knowledge in decisions
Fairness	<ul style="list-style-type: none"> Procedural fairness Fairness of grading process Fairness of results Minimizing grading bias Minimizing human error Appropriate evaluation of work
Organizational oversight	<ul style="list-style-type: none"> Responsible person Accountability
Transparency	<ul style="list-style-type: none"> Avoid black-box
Autonomy	<ul style="list-style-type: none"> Right to not use AI
Responsibility	<ul style="list-style-type: none"> Employ experts with domain knowledge Responsibility over teaching content Appropriate AI model use Responsibility to protect learners
Soft skills	<ul style="list-style-type: none"> Social skills for educators Deliberate and reflective AI use Goal-oriented AI use
Human-decision making	
Academic honesty	<ul style="list-style-type: none"> Disclosure of AI use
Data privacy	
Critical thinking	
Informational literacy	
Awareness	<ul style="list-style-type: none"> Awareness of limitations of AI Awareness of AI bias Awareness of potential unreliability
Teaching competencies	<ul style="list-style-type: none"> Soft skills Technological awareness AI literacy
Support	<ul style="list-style-type: none"> Institutional support in AI use Educator training in AI Continuous training in AI
Digital equity	<ul style="list-style-type: none"> Reduce digital divide for learners AI system cost accessibility
Trust	<ul style="list-style-type: none"> Impartiality
Creativity	<ul style="list-style-type: none"> Role of human creativity

	Practical AI literacy Theoretical AI literacy Awareness of AI ethics Recognize AI content Awareness of AI types Knowledge of history of AI Deep awareness of AI functions Knowledge of epistemology Machine learning competencies Awareness of AI lifecycle Data literacy
AI literacy	
Augmented human abilities	AI as a collaborative partner
Truth	
Critical thinking	
Collaboration	Stakeholder collaboration Cross-disciplinary collaboration
Integration of AI into professional training	
Prevent surveillance	
Prevent unfair discrimination	
Explainability	

7.3 Analysis of main themes from empirical literature

7.3.1 Academic dishonesty & fairness

Preventing academic dishonesty is a central theme within the literature on AI and education (Adžić et al., 2024; del Álamo Cienfuegos et al., 2024; Leoste et al., 2025; Pavlova & Slavov, 2025; Söderström et al., 2024). Defining what is fair use of AI and what counts as academic dishonesty is a concern for both educators and learners. For instance, educators may agree that submitting AI-generated work as original student work is unacceptable and risks committing plagiarism (Pavlova & Slavov, 2025).

However, it may seem unrealistic to completely prohibit the use of commercial available AI-enabled tools to learners, as well as being counter-intuitive for fostering AI literacy (Prather et al., 2023). The work of defining fairness for AI-use in education is more often than not, landing on the discretion of individual institutions and instructors who lack support and guidance in addressing AI use in school work (Prather et al., 2023; Söderström et al., 2024).

For students, there may be uncertainty about how to approach using AI in their work. In many institutions, consequences for not disclosing AI use or misusing AI can be serious and result in academic sanctions for plagiarism (Adžić et al., 2024; Pavlova & Slavov, 2025). On the other hand, there is often a lack of established guidelines on what can be considered fair use of AI by learners causing confusion and unintentional misuse (Söderström et al., 2024).

7.3.2 Automated grading systems & fairness

Unfair grading bias from humans towards certain students can be potentially avoided with the use of AI grading and teaching tools (Figueras et al., 2024). AGS's could therefore limit the bias associated with human grading and ensure fairer results (Figueras et al., 2024). Students mention that bias towards non-native speakers and diverse cultural backgrounds in particular may be avoided by the use of AI grading tools that focus on content and accuracy rather than form (Figueras et al., 2024).

However, the performance capabilities of AGS's should be such that they are able to assess more nuanced aspects of students' performance rather such as the thought process and effort put into answers (Figueras et al., 2024). Furthermore, there is the risk of inherent biases from training datasets against gender, language, ethnicity and other protected characteristics affecting the assessment of student work (Figueras et al., 2024).

Educators are also concerned about AGS's having sufficient privacy protection for student data (Figueras et al., 2024). Misuse of data from education is a major concern for fairness because of the potential for leaked student work to be used for academic misconduct and cheating or outsourced to third parties for unintended uses (Figueras et al., 2024).

Although the ability to have human oversight with AGS's can present a solution to some situations, human involvement can also be subject to unfair bias as well as data security concerns. Thus, the deployment of AGS's should be approached cautiously and with the

appropriate safeguards and evaluation tools to ensure that grading is consistent with educational guidelines and fair.

7.3.3 Access to benefits of AI tools

Additionally, being able to use powerful AI tools may also be contingent on access to finances and institutions with the appropriate infrastructure and support for AI use (Pavlova & Slavov, 2025; Prather et al., 2023). There is the risk that learners with lower-income may not have access to higher paid tiers of commercially available AI tools which have paid tiers (Pavlova & Slavov, 2025). Institutions without adequate support to implement AI-enabled technologies or support the training of students and educators are also disadvantaged. Educational AI systems may not cater to learners who are in the minority (Bucea-Manea-Țoniș et al., 2022).

Another issue is the performance inconsistencies of currently commercially available LLMs which may make them less useful for non-English languages (Cernadas & Fernández-Delgado, 2021). While English may be used as the common language in many settings, the EU is immensely diverse in the languages used amongst member states. Poor performance in non-English languages in LLMs may also mean that the potential benefits of using LLMs is unevenly distributed amongst those who can speak English, and those who have a lesser command of it.

7.3.4 Changes to learner and educator skills

Overreliance on AI use may detrimentally affect learners' critical thinking and creativity. Educators have stated in studies that students risk affecting their critical thinking skills if they rely on LLMs to reason through problems for them (Adžić et al., 2024; Figueras et al., 2024; Pavlova & Slavov, 2025). Relying on LLMs to generate art or text may also affect students' creativity, depending on the context (Pavlova & Slavov, 2025). On a broader level, LLM overreliance can also lead to passive learning and degradation of learning skills in general by offering a way for learners to offload cognitive labour (Adžić et al., 2024; Tenório & Romeike,

2024). Therefore, students using LLM models to “shortcut” achievement of learning objectives will be encouraged to value outcomes over the learning process (Figueras et al., 2024).

However, the use of LLMs and other AI-enabled technologies (such as including GenAI) are not necessarily definite risks to important learner competencies. They can also stimulate critical thinking skills and creativity through augmenting human abilities, saving time and enhancing the educational experience (Pavlova & Slavov, 2025; Prather et al., 2023).

Many of the concerns about skill change and loss are tied to calls to adjust pedagogical and assessment strategies to the potential use of LLMs by students (Prather et al., 2023; Kalving et al., 2024). Some teaching professionals even state a desire to re-evaluate education systems to centre learning based on critical thinking, creativity and demonstrated understanding of the content instead of other milestones such as memorization of content (Prather et al., 2023).

... if you add a good programmer to large language models, you get two good programmers...if you add a mediocre programmer to large language models, you get just large language models. So, from that perspective... if you want to really be able to amplify the capabilities of humans, we need to make sure that the basic competencies remain. (Prather et al., 2023, p. 25)

On the other hand, there are concerns that the use of LLMs will change the role of teachers unfavourably. For instance, the teacher-student relationship could drift towards a more supervisory role over AI systems which dictate learning processes rather than an emphasis on connection with students (Figueras et al., 2024). It is even a possibility that educators will pass responsibility of student learning off to AI systems, shirking accountability of grading and teaching errors (Figueras et al., 2024). Educators also express a worry that their jobs are endangered by the introduction of AI-enabled learning technologies (Klemettilä et al., 2025).

However, it is still unclear as to whether AI systems will reach a level which could be worrisome to the teaching job market or the role of a teacher (Figueras et al., 2024; Klemettilä et al., 2025; Pavlova & Slavov, 2025).

Therefore, there appears to be a necessity to shift priorities on what skills are important for educators and learners to have. Instead, the focus on soft skills such as critical thinking, creativity and communication may become more emphasized in the future as the dynamics within the classroom change (Jemetz et al., 2024).

7.3.5 Keeping education a neutral space

Research has shown that some educators worry about the influence of government and corporations on students through LLMs which may promote a biased agenda. This possibility threatens the value of truth and impartiality in academia. The power that governments, corporations and other actors must exert control over the type and amount of knowledge that students can access may present a threat to content credibility as well as manipulation of information (del Álamo Cienfuegos et al., 2024; Klemettilä et al., 2025). Governments may exert their influence via propaganda or state interests while on the contrary, the centralization of power into a few technological companies could impact the dissemination of information to learners (del Álamo Cienfuegos et al., 2024).

Potential solutions include multistakeholder involvement and oversight onto educational AI-enabled technologies, mixed funding sources and improving digital and AI literacy for educators and students (del Álamo Cienfuegos et al., 2024; Klemettilä et al., 2025). Overall, transparency, accountability, oversight and AI governance were stated as key to preventing partiality in the use of AI systems in education.

Figure 11

Opportunities for the use of AI in education, extracted from the selection of empirical literature

Category of use	Sub-uses
Utility of AI for educators	Personalization of learning content development Reduce repetitive tasks Save time Summarize large amounts of data Summarize texts Arrive at quick results Grading support Personalized teaching materials
Utility of AI for students	Multilingual support Personalized learning content Reducing barriers of entry to new fields (such as computer science) Improve performance

Figure 12

Relevant trade-offs present in the empirical literature

Benefits	Trade-offs
Utility Accessibility to information/efficiency	Limited diversity of information Echo chamber Information manipulation Misinformation
Transparency Explicability of AI-enabled grading processes	Gaming the system Academic dishonesty
Personalization Adaptability	Impractical to do for every student
Convenience Automation	“Laziness” Risk to quality of output Risk to creativity

Augmentation of human skills Amplify abilities of skilled (users of AI)	Degradation of skills for less-skilled or of low-digital literacy/misuse/risk to learning outcomes
---	--

Figure 13

Vulnerable populations in education as identified in the empirical literature

Vulnerable population	Explanation
Low-income students and learners in lower resource institutions	Unequal access to helpful AI technologies Unequal access to digital literacy training
Women and girls	Risk of low digital literacy (discrimination in STEM fields)
Minorities and underrepresented groups	Unequal access to AI Potential of grading bias in automated grading systems Risk of gender bias in datasets Risk of ethnicity bias in datasets Risk of bias towards non-English languages
Older adults	Risk of digital exclusion based on skills
non-English speakers	Risk of low LLM performance
Students with diverse learning needs	Educational AI systems may not cater to learners who are in the minority
Students and teachers with low digital literacy	Risk of misuse of AI Risk of overreliance on AI Risk of digital exclusion
Teachers (job security)	Risk of job replacement by AI
Learners in fields with high exposure to AI (job security)	Risk of job replacement by AI Risk of insecure career prospects

7.4 References for education related papers

Adžić, S., Savic-Tot, S., Vukovic, V., Radanov, P., & Avakumović, J. (2024). Understanding student attitudes toward GenAI tools: A comparative study of Serbia and Austria. *International Journal of Cognitive Research in Science, Engineering and Education*, 12(3), 123–145. <https://doi.org/10.23947/2334-8496-2024-12-3-583-611>

- Bucea-Manea-Țoniș, R., Kuleto, V., Gudei, S. C. D., Lianu, C., Lianu, C., Ilić, M. P., & Paun, D. (2022). Artificial intelligence potential in higher education institutions enhanced learning environment in Romania and Serbia. *Sustainability*, 14(10).
<https://doi.org/10.3390/su14105842>
- Cernadas, E., & Fernández-Delgado, M. (2021). Embedded ethics to teach machine learning courses: An experience. *2021 XI International Conference on Virtual Campus (JICV)*, 1–4. <https://doi.org/10.1109/JICV53222.2021.9600426>
- del Álamo Cienfuegos, A., Lis, P., Zadorozhna, O., Espiritusanto, O., Gawronska-Nowak, B., & Zarzycka, A. (2024). AI: What do we fear? What do we hope for? Perception of the societal impact of AI in a European transnational cross-sectional study. *2024 IEEE International Conference on Engineering, Technology, and Innovation (ICE/ITMC)*, 1–9. <https://doi.org/10.1109/ICE/ITMC61926.2024.10794321>
- Figueras, C., Rossitto, C., & Cerratto Pargman, T. (2024). Doing responsibilities with automated grading systems: An empirical multi-stakeholder exploration. *Proceedings of the 13th Nordic Conference on Human-Computer Interaction, NordiCHI '24*.
<https://doi.org/10.1145/3679318.3685334>
- Jemetz, M., Dolezal, D., & Motschnig, R. (2024). Secondary teachers' self-perceived AI competences in relation to renowned European digital competence frameworks. *Informatics in Schools: Innovative Approaches to Computer Science Teaching and Learning, ISSEP 2024, Lecture Notes in Computer Science, 15228*, 3–17.
https://doi.org/10.1007/978-3-031-73474-8_1
- Kalving, M., Colley, A., & Häkkinen, J. (2024). Where AI and Design Meet—Designers' Perceptions of AI Tools. *Proceedings of the 13th Nordic Conference on Human-Computer Interaction, NordiCHI '24*. <https://doi.org/10.1145/3679318.3685389>
- Klemettilä, P. A., Sharma, S., Mochiyama, F., Iivari, N., Iwata, M., & Koivisto, J. (2025). “It’s just a machine that predicts”—Demystifying artificial intelligence/machine learning with teenagers. *Proceedings of the 24th Interaction Design and Children, IDC '25*, 168–182.
<https://doi.org/10.1145/3713043.3728853>
- Leoste, J., Pöial, J., Kivisalu, E., Marjanovic, U., Rakic, S., & Robal, T. (2025). Integration of artificial intelligence in higher education programming courses: Insights from student perspectives and practices. *Lecture Notes in Networks and Systems, 1260*, 417–427.
https://doi.org/10.1007/978-3-031-85652-5_42
- Pavlova, Y., & Slavov, V. (2025). Exploring the ethical use of AI technologies/applications in academic environments (M. Kostov & M. Atanasovski, Eds.). *2025 60th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)*, 1–4. <https://doi.org/10.1109/ICEST66328.2025.11098253>
- Prather, J., Denny, P., Leinonen, J., Becker, B. A., Albluwi, I., Craig, M., Keuning, H., Kiesler, N., Kohn, T., Luxton-Reilly, A., MacNeil, S., Petersen, A., Pettit, R., Reeves, B. N., &



HORIZON-MSCA-2023-DN-01

Grant Number 101169473



Funded by the
European Union

Savelka, J. (2023). The robots are here: Navigating the generative AI revolution in computing education. *Proceedings of the 2023 Working Group Reports on Innovation and Technology in Computer Science Education, ITiCSE-WGR '23*, 108–159.

<https://doi.org/10.1145/3623762.3633499>

Söderström, U., Hedström, E., Lambertsson, K., & Mejtoft, T. (2024). ChatGPT in education: Teachers' and students' views. *Proceedings of the European Conference on Cognitive Ergonomics 2024, ECCE '24*. <https://doi.org/10.1145/3673805.3673828>

Tenório, K., & Romeike, R. (2024). AI Competencies for non-computer science students in undergraduate education: Towards a competency framework— *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research, Koli Calling '23*. <https://doi.org/10.1145/3631802.3631830>

8. Healthcare and mental health

A total of 3/50 articles in the sample focused on healthcare as a topic with none about mental health specifically.

After conducting a scan of Google Scholar and Perplexity.ai using relevant keywords from the initial search strategy, one relevant report containing empirical evidence and published between 2018-2025 was located. A cursory search revealed no relevant academic articles. As no EU-specific values were able to be extracted, insights from the report were combined into other normative values from soft policy and normative documents.

These countries were represented: Spain (n=1), Poland (n=1), and the Netherlands (n=2). In terms of methodology, 2/3 were qualitative and 1/3 was mixed-methods. 1/3 articles had participants from more than one country. The total number of participants in the three articles was 121.

For the purposes of this analysis, healthcare is used as a proxy for mental health.

8.1 Overview

Whether AI-enabled tools are useful for practitioners and patients is paramount to their ethical implementation (World Health Organization, 2024). Promoting patient wellbeing and avoiding harm through trustworthy AI-enabled tools is a stringent requirement.

In the EU, on-the-ground expertise is likely necessary to be able to implement systems that address the needs of local contexts as well as overarching regulations and laws. For instance, research has shown that European contexts place particular emphasis on the ability for tools to integrate into existing clinical workflows and to be adaptable to different situations (Maris et al., 2024). This is perhaps related to the diversity of cultural and linguistic contexts that exist within the EU as well as medical systems that must all meet harmonized regulations such as the

GDPR and the EU AI Act (Maris et al., 2024). Tools that enter the EU market must therefore be EU-compliant as well as useful for a range of languages and cultural settings to be successfully integrated into the clinical landscape. In all cases, the safety and effectiveness of AI systems is critical.

Furthermore, practitioner-patient dynamics inherently involve vulnerability on behalf of patients who are dependent on the expertise and judgment of medical professionals. Human oversight of AI systems across all uses appears to be important, with trust in devices closely intertwined with trust in the medical expertise of responsible practitioners (del Álamo Cienfuegos et al., 2024). However, changes in the capabilities of AI systems may also have the possibility to change requirements for human oversight as performance improves along with patient outcomes.

In sum, identified values in the literature point to curiosity about the use of AI to further patient and practitioner wellbeing, as well as apprehensiveness about possible risks to patient wellbeing and practitioner jobs.

8.2 Normative values in healthcare and AI

In terms of ethical guidance, the field of bioethics can provide established principles, such as Beauchamp's *autonomy*, *nonmaleficence*, *beneficence* and *justice* (Beauchamp & Childress, 1979). These principles were formulated largely to guide the behaviour of medical practitioners and researchers towards human subjects (Beauchamp & Childress, 1979). The work of interpreting them in light of the integration of AI systems introduces new challenges such as defining the behaviour of AI systems towards patients and practitioners and the values they should embody (Figure 14). Sub-principles such as *transparency* and *trust* are necessary to define the instrumental goals that lead to the fulfillment of more abstract goals such as *beneficence*, as well as the values that AI systems in general should align to (Figure 14).

Figure 14

Normative values on AI and healthcare from grey literature

Function	Value	Description
Instrumental societal goods to promote/contribute/uphold	Informed consent	Ensure patients are fully informed and understand AI systems, limitations, and benefits ¹
	Informed design	Align AI design with ethical and regulatory requirements ¹
	Quality control	Evaluate and improve AI system quality over time ¹
	Stakeholder consultation	Meaningfully consult the public and stakeholders on AI design and use ¹
	Evaluation	Enable evaluation of AI systems by patients, practitioners, and impacted stakeholders ¹
	Human supervision	Incorporate human oversight throughout the AI lifecycle ¹
	Redress	Provide mechanisms for individuals and groups adversely affected by AI decisions to seek redress ¹
	Monitoring	Continuously monitor AI systems to identify disproportionately harmful effects on groups or individuals ¹
	Professional standards	Ensure AI is used by appropriately trained individuals under proper conditions ¹³⁴
	Global digital equity	Ensure AI is implemented effectively promoting capacity building and autonomy of local communities ¹
	Human rights and dignity	Uphold dignity and avoid harmful bias toward marginalised groups based on personal characteristics ³⁴
	Therapeutic ethical standards	Uphold expected and applicable therapeutic ethical standards when employing AI systems ³
	Healthcare system sustainability	Promote long-term sustainability and resilience of health systems for future generations ¹
	Environmental sustainability	Limit environmental harm from AI use and promote environmentally sustainable practices ¹
Inclusiveness	Ensure healthcare access irrespective of personal characteristics ¹	
Cultivate human autonomy	Avoid overreliance on AI and recognise potential risks on human autonomy ¹	

	Foster accessible health systems	Avoid limiting affordability and accessibility of healthcare due to inequitable access to AI ¹
	Counter systemic bias	Prevent embedded biases in datasets from perpetuating societal inequities and negative impacts on patients and providers ¹
	Respect for human labour	Avoid negative labour impacts including job loss, job quality, and losing skilled workers ¹
	Ethical governance	Maintain governance structures and ethics committees for AI oversight ⁴
	Accountability	Ensure accountability measures are put in place for AI systems and practitioners ³
Desired qualities and behaviours for human beings to have in regard to AI	Non-maleficence	Avoid AI use when risks of harm are present and alternatives exist ¹²³
	Adherence to professional standards	Ensure AI is used only by trained personal under appropriate professional conditions ¹
	Fidelity and responsibility	Uphold professional duties and cooperate with others in healthcare system to serve best interests of patients and relevant parties ³
	Ensure cybersecurity	Prevent malicious attacks undermining the confidentiality of patient data and safety and trust of patients in healthcare system ¹
	Beneficence	Promote welfare of patients and relevant parties ²³
	Lawfulness	Abide by all relevant laws and regulations concerning but not limited to data privacy and patient and practitioner relationships ³⁴
	Maintain confidentiality	Abide by all relevant laws and regulations concerning but not limited to data privacy and patient and practitioner relationships ⁴
	Ethical awareness	Remain aware of ethical codes and standards when engaging with AI systems ⁴
	Integrity	Commit to promoting honesty, accuracy and truthfulness in all activities ³
Desired qualities for AI to have	Lawful	Ensure that AI systems behave according to appropriate legal frameworks for data protection, patient privacy and safety, not delivering

		inappropriate healthcare outside of the health system, or inappropriately to unpermitted patient groups (e.g. children) ¹
	Ethical	Adheres to bioethical principles and ethical principles ¹
	Epistemically just	Respecting the lived experiences of patients as knowledge-holders ¹
	Equitable	Avoiding harmful bias against marginalised groups in particular ¹
	Inclusive	Ensure access and use of healthcare is available to all patients irrespective of personal qualities or other characteristics ¹
	Non maleficent (AI systems)	Safeguarding welfare to prevent harm against patients, providers, and other relevant parties ²³
	Beneficent (AI systems)	Promoting welfare in patients and providers ²³
	Continuing responsibilities	Maintaining ethical standards and confidentiality after engagement with patient ⁴
	Meta-cognitive awareness	Avoiding conflicts of interest, manipulation and misuse of power in patient engagements ⁴
	Personalized care	Supporting patient empowerment, education, autonomy and quality of care through personalized treatment ¹
Behaviours and qualities that AI systems should avoid	Avoid inaccurate statements	Prevent false or misleading AI outputs causing harm ¹
	Avoid manipulation	Avoid manipulative behaviours that harm patient wellbeing ¹
	Avoid infringement of privacy	Protect patient health information and confidentiality ¹
	Avoid impact on fiduciary relationship	Protect the clinician–patient relationship ¹
	Avoid epistemic injustice	Avoid undermining patients' knowledge and lived experiences or preventing them from being believed or informing their own care ¹
	Avoid inappropriate delivery outside system	Avoid healthcare delivery outside of scope of care and established professional boundaries ¹

	Avoid inappropriate healthcare delivery to children	Prevent harm to children and ensure age-appropriate safeguards ¹
	Avoid harmful bias against marginalized groups	Prevent discrimination against marginalised groups and perpetuating existing societal injustices ¹
Potential risks to the health system	Overreliance on AI	Excessively relying on AI systems to provide solutions ¹
	Inaccessibility	Limiting the affordability and accessibility of healthcare due to use of LLMs that pose financial, digital, technical and other barriers ¹
	Systemic bias	Presence of harmful bias that impacts healthcare systems due to embedded biases in large datasets ¹
	Labour impacts	Negative impacts including loss of skilled workers, job losses and decreased quality of care ¹
	Dependence on ill-suited LLMs	Perpetuation of the digital divide via poor implementation and maintenance of systems not fitted to contexts ¹
	Cybersecurity risk	Risk of bad actors and attacks on healthcare system compromising patient data and system functionality ¹
Values to consider upon deployment of AI in healthcare	Usefulness, usability and efficacy	Beneficial, reliable AI solutions that are usable and fit into existing workflows ⁵
	Fairness, equity and bias management	Fair AI solutions that work equally well for all demographic groups and have adequate bias management ⁵
	Safety and reliability	Non-harmful AI solutions that have been adequately tested and evaluated, and are held to accountability and governance structures ⁵
	Transparency, intelligibility and liability	Awareness of AI's capabilities and limitations amongst stakeholders and identifying liable individuals in case of harm ⁵
	Security and privacy	AI systems that protect data confidentiality and integrity in compliance with regulations and ongoing monitoring ⁵
	Health equity and wellbeing	Ensure fair and accessible healthcare use and access for all groups,

		especially disadvantaged and marginalised populations ¹
	Human autonomy	Ensure humans have control over medical decision-making and healthcare systems ¹
	Privacy	Protect the privacy and confidentiality of patients and providers ¹
	Justice	Strive for fairness and equitable distribution of benefits and burdens in healthcare ²

Note. Based on a thematic extraction of

- (1) World Health Organization. (2024). Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models. <https://www.who.int/publications/i/item/9789240084759>
- (2) Beauchamp, T. L., & Childress, J. F. (1979). Principles of biomedical ethics. Oxford University Press.
- (3) American Psychological Association. (2017). Ethical principles of psychologists and code of conduct (2002, amended effective June 1, 2010, and January 1, 2017). <https://www.apa.org/ethics/code/>
- (4) European Federation of Psychologists' Associations (EFPA). (2025). *Meta-Code of Ethics*. https://www.efpa.eu/sites/default/files/2025-06/2025-06-03_efpa-meta-code-of-ethics.pdf
- (5) Gooding, P., van Kolfschooter, H., & Centola, F. (2024). Artificial Intelligence in mental healthcare. Mental Health Europe. <https://easped.eu/news-detail/ensuring-a-responsible-use-of-artificial-intelligence-in-mental-healthcare/>

Normative values extracted from the empirical literature place more focus on the role of AI in benefitting patient empowerment/autonomy and how AI-enabled tools can be implemented in accordance with promoting patient wellbeing (Figure 15). For instance, characterizing *Transparency* with sub-values such as the “*Right to human explanation about AI decisions*” provides useful information about the concerns of patients and how values can actionably promote their preferences.

Figure 15

Normative values extracted from empirical literature on healthcare

Normative values extracted from articles	Sub-values
AI supporting patient autonomy	Shared decision-making Informed consent AI strengthened decision making Dignified care Mitigate potential unfair treatment by human practitioners Assist patient advocacy Assist patient empowerment
AI as practitioner support	AI-augmented human capabilities and judgment Second opinion Collaborative decision making Brainstorming
Responsibility and accountability	Patient and practitioner in charge of final decision Evaluation Reporting
Governance	Robust rules and regulations Ethical guidelines Adhere to medical ethics Data regulations Harmonized standards for AI tools Follow established consensus methodologies
Human oversight	Practitioner oversight of AI tools Practitioner as mediator Human quality control
Usability	Integrated AI into existing workflow Adaptable Personalized to patient needs Integrate diagnostic, prognostic, monitoring functionalities Flexible and adaptive
Universality	Interoperability of medical systems Integration of different types of medical data Applicability across diverse populations Applicability across different countries (EU) Comprehensiveness Reproducibility of results Universal standards
Trust	Trust in practitioners to explain AI decisions Trust in practitioners to intervene Trust in AI developers
Evidence-supported AI use	AI strengthened decision making Randomized control trial supported AI validation and evaluation Data-driven AI use

Robustness	Safety Cybersecurity Accuracy of clinical decisions Reliability Data diversity in training
Right to a human	Human (practitioner) responsibility of care Human (practitioner) expertise Human touch Preserve the fiduciary relationship
Ethical research	Consent of data use (for research) Separation between research data consent and treatment consent
Fairness	Non-discriminatory treatment Understandability of AI system (to patients with low digital literacy) Patient representation in AI development Procedural fairness of AI decisions AI as neutral party
Transparency	Right to human explanation about AI-made decisions Explicability of AI decisions Informed patient decision-making Transparent track record of AI tool available to patient
Sustainability	Environmental sustainability Supporting sustainable practices AI-supported resource usage Avoid medical resource wastage System sustainability Improve financial sustainability of treatments Improve human resource capacity Improve medical professional capacities
Societal progress	Medical advancements Augmented societal capacities Technological innovations Improved patient outcomes

8.3 Analysis of main themes from empirical literature

8.3.1 Human oversight

Involving practitioners and professionals in decision-making when using AI systems in healthcare appears to be a keystone for responsible AI use. The desired level of human

oversight, however, can differ depending on various factors such as the type of patient-engagement and the reliability of the system (Maris et al., 2024).

Guidelines and the literature also mention the necessity for trained medical professionals to supervise AI-assisted decision-making processes to prevent errors and act as a responsible party (Maris et al., 2024; Puig-Bosch et al., 2024; World Health Organization, 2024). The implication is that humans should have ultimate decision-making power in the healthcare sector and that human medical decision-making may be more conducive to respecting patient autonomy. Overall, one aspect of human supervision as seen in the literature is closely tied to the primacy of human care in medicine as well as the authority of the medical practitioner in medical decision-making.

Impacts on the patient-practitioner relationship

However, there is often an inherent power dynamic in the fiduciary relationship between the patient and practitioner which puts the former in a vulnerable position. Classical bioethics emphasizes beneficence, promoting the wellbeing of patients, and nonmaleficence, “do no harm”, as key principles for practitioners to abide by (Beauchamp & Childress, 1979). AI systems may not be able to be bound to the same obligations as practitioners but can be tied to responsible figures who can be accountable for decisions and outcomes.

AI and patient autonomy

As well as being used as tools for practitioners, AI systems can also be utilized by patients to support their autonomy in healthcare encounters. Given the possibility that practitioners make errors or hold harmful biases against patients with certain characteristics, having an AI system provide an opinion as a third party is a desirable option (Maris et al., 2024). For instance, patients imagine utilizing AI-enabled tools as neutral parties and sources of additional

D3.1 Impact Assessment of Values Affected

information when presented with medical decisions (Maris et al., 2024). Furthermore, AI systems may help patients make sense of medical data to provide actionable insights that can help inform their decision-making (del Álamo Cienfuegos et al., 2024). Patients mention that AI-enabled tools could significantly improve access to health information and self-treatment in the case of disabled individuals and older adults (del Álamo Cienfuegos et al., 2024). In this regard, patients may see access to more AI-enabled tools as a way inform and protect themselves when faced with complex medical situations (del Álamo Cienfuegos et al., 2024; Maris et al., 2024).

In many senses, AI is seen as a supportive tool for patient empowerment or practitioner decision-making, but not an entity capable of having responsibility over decisions. Patients may desire an accountable party for AI-assisted decisions and find it unacceptable for AI systems to have the power of making the final decision on their care (Maris et al., 2024). This also relates to human autonomy and retaining the right to a human practitioner, as well as the desire for a human practitioner-patient relationship (Puig-Bosch et al., 2024).

AI and mitigating medical bias

Other marginalized groups in society could also benefit from increased accessibility to traditionally gate-kept forms of medical knowledge and treatment that they encounter socio-economic or systemic barriers to. Thus, AI can be used to subvert the traditional fiduciary power dynamic and allow patients to have agency in their interactions with medical systems. The reverse can also be true, where human oversight over AI systems is necessary to mitigate harms such as unfair algorithmic bias and embedded stereotypes within training data of LLMs (World Health Organization, 2024).

Additionally, one end goal of involving human oversight throughout the AI lifecycle in medicine is to allow people to have agency over important decisions which impact their health and others. The idea of an AI system operating without supervision is unacceptable to many in high-stakes situations such as in healthcare (Maris et al., 2024; Puig-Bosch et al., 2024). While this sentiment may change as AI systems develop and become increasingly more capable and reliable, a core tenet of oversight is still to ensure that both practitioners and patients have control over what happens in a healthcare setting.

8.3.2 Data privacy

The trade-off between the use of health information to improve the performance of AI systems, and the potential risks to data privacy and security is a prominent theme in the literature.

Patients and practitioners express optimism about the use of AI systems to power technological advancements such as personalized medicine, predictive tools and precision medicine (del Álamo Cienfuegos et al., 2024; Maris et al., 2024). These hopes are closely accompanied by the fact that the use of health-related data when integrating AI systems into healthcare settings may risk data privacy and security. However, the progress to be made regarding medical treatments and patient health outcomes seems to be a major driver of continued interest and desire for AI system use in healthcare (Puig-Bosch et al., 2024).

Preconditions for the implementation of AI systems into healthcare mentioned in ethical guidelines are in large part geared towards data regulations and standards specifically shaped to the privacy requirements for using sensitive health information, strong cybersecurity and adherence to laws by practitioners and institutions (World Health Organization, 2024). These are in part shaped by standard healthcare expectations such as to maintain confidentiality and ensure that informed consent to disclose patient information is garnered (American Psychological Association, 2017; Beauchamp & Childress, 1979; European Federation of

Psychologists Associations (EFPA), 2025). While these original principles are maintained with the introduction of AI systems in healthcare, the scale of which data is used and collected has increased multifold, creating new challenges.

Ethics around using health data overlap with legalistic elements such as fulfilling institution and state-level standards, as well as the General Data Protection Regulation (GDPR) in Europe (Puig-Bosch et al., 2024). These existing protections against unethical use of health data will potentially exist in addition to those put into place with the full implementation of the EU AI Act (2024). To comply with standards, healthcare AI-enabled tools may need to adhere to rules such as continuous monitoring and documentation by deployers (Regulation (EU) 2024/1689, 2024, Art. 26) and or inform users when they are interacting with an AI system (Regulation (EU) 2024/1689, 2024, Art. 50). The protections of these largely map onto patient and practitioner concerns expressed in the literature which surround informed consent about type of data use by AI systems (research vs. medical), data security and robustness (Maris et al., 2024; Puig-Bosch et al., 2024).

8.3.3 Practitioner-patient relationship and “the human touch”

Patients express a desire to preserve a relationship with a human practitioner in the midst of an increased interest in AI use in the healthcare system (Maris et al., 2024; Puig-Bosch et al., 2024). Many of these sentiments in the literature were related to patient integrity and being treated like an individual, which was felt to be more of a possibility with a human practitioner (Maris et al., 2024).

...most participants did not expect that using AI would necessarily lead to significantly different diagnostic and treatment outcomes either, and in that respect, held similar expectations for both AI and doctors. However, many participants were concerned that an AI-driven decision would reduce the patient to a disease or a set of somatic symptoms, thus compromising human integrity and respect for one's individuality (Maris et al., 2024, p. 8)

It was also acknowledged by patients that the time constraints and resource constraints faced by practitioner could also impact their feeling of being valued as individuals (Maris et al., 2024). Additionally, "the human touch" may overlap conceptually with human oversight and quality control, implying a belief that human decision-making may outperform AI system decision-making in certain cases (Puig-Bosch et al., 2024). While there are also other worries that human practitioner roles could be replaced by AI automated systems, it appears that core elements to care still require a human presence for important decisions at minimum.

8.3.4 Transparency and trust

Furthermore, the transparency of outputs from AI-enabled tools is paramount in a medical setting. *Transparency* is a value related to a number of other important considerations and values in the literature such as *trust*, *explainability* and *informed consent*. In the context of healthcare, it is often associated with how and why decisions come to be made. Trust in the tool, the practitioner and/or the healthcare system can be affected by whether patients feel that decisions about their health are adequately transparent to them (Puig-Bosch et al., 2024). Lack of transparency may pose a risk to patients' informed consent as well as practitioners' ability to harness data-driven insights. As medical decisions can feel complex and opaque to the layperson, an additional layer of complexity imposed by AI algorithms may make it more difficult

for both practitioners and patients to understand the rationale behind outputs (Puig-Bosch et al., 2024).

Patients are concerned about transparency with AI algorithms no longer set by humans, which complicates doctors' ability to explain decisions to patients" (Puig-Bosch et al., 2024, p. 4)

Additionally, explanations that are less acceptable could in turn put more pressure into the requirement for human-oversight as patients rely more on practitioners' expertise to decipher the output (Maris et al., 2024). A system lacking transparency could therefore impact their usefulness in the medical space.

Figure 16

Opportunities for the use of AI in healthcare identified in the empirical literature

Instrumental uses for AI in healthcare	Sub-uses
Support patient empowerment	<ul style="list-style-type: none"> Access to medical data information Provide second opinion as neutral party Mitigate medical discrimination and unfair bias Personalized treatment Explanation of complex medical information
Support clinical work	<ul style="list-style-type: none"> Precision medicine Harness large amounts of medical data Assistance with decision-making Access to medical information more readily Second opinion on decisions
Support healthcare system	<ul style="list-style-type: none"> Supplement human resource shortages Predict resource use Prevent excessive medical waste

Figure 17

Trade-offs between values in healthcare identified in the empirical literature

Benefits	Trade-offs
<p>Patient empowerment/Patient independence</p> <p>Accessibility to second opinions Accessibility to medical information Greater accessibility to treatment</p>	<p>Patient wellbeing</p> <p>Chance of error Inaccurate information Hallucinated information Risk of harmful or offensive outputs Misinterpretation of presented information by patient causing harm</p>
<p>Better healthcare system sustainability</p> <p>Saving time Reduce need for human resources (amid staff shortages and budget constraints)</p>	<p>Practitioner wellbeing</p> <p>Risk of loss of skilled professionals Risk of skill loss Risk of job replacement Risks from inconsistent regulations and guidelines on AI in healthcare</p>
<p>Enhanced practitioner abilities</p> <p>Increase capacity to take care of patients by reducing task load Mitigate errors and mistakes Brainstorming tool Second opinion Access to medical information Access to insights from medical data Reduce cognitive workload</p>	<p>Fiduciary relationship</p> <p>Risk of over-reliance on AI Risk of losing human connection to patient Risk of important skill loss Risk of lack of human oversight on AI-supported decisions Loss of quality of care</p>
<p>Data-driven medicine</p> <p>Harness large amounts of medical data Integrate and analyze different types of medical data Accessible data-driven insights available to practitioners and patients Support data-driven decision making</p>	<p>Privacy and transparency</p> <p>Risks to privacy and confidentiality of sensitive data Risk of cybersecurity attacks Risk of data misuse Risk of data overwhelm for patients and practitioners Lack of transparency of decisions Lack of explainability of decisions Data opaqueness Black-box effect</p>

Figure 18

Vulnerable populations in healthcare

Vulnerable population	Explanation
Medical professionals	Risk of labour changes (especially older practitioners) Risk of skill devaluation in society Risk of skill degradation Risk of overreliance on AI
Disabled individuals	Risk of overreliance on AI Risk of inaccurate information from AI Replacement of human touch with AI
Individuals from developing countries	Risk of unequal access to AI and treatment Risk to global health equity
Individuals with less digital literacy	Risk of unequal access to AI and treatment Poor understandability of AI and less access to benefits
Older adults	Risk of poor digital literacy Digital barriers to care Replacement of human touch with AI
Children	Risk of inappropriate healthcare provision outside scope of AI system
Patients (generally)	Risk to patient data privacy Risk of algorithmic discrimination to marginalized groups

8.4 References for healthcare related papers

- del Álamo Cienfuegos, A., Lis, P., Zadorozhna, O., Espiritusanto, O., Gawronska-Nowak, B., & Zarzycka, A. (2024). AI: What do we fear? What do we hope for? Perception of the societal impact of AI in a European transnational cross-sectional study. *2024 IEEE International Conference on Engineering, Technology, and Innovation (ICE/ITMC)*, 1–9. <https://doi.org/10.1109/ICE/ITMC61926.2024.10794321>
- Maris, M. T., Koçar, A., Willems, D. L., Pols, J., Tan, H. L., Lindinger, G. L., & Bak, M. A. R. (2024). Ethical use of artificial intelligence to prevent sudden cardiac death: An interview study of patient perspectives. *BMC Medical Ethics*, *25*(1). <https://doi.org/10.1186/s12910-024-01042-y>
- Puig-Bosch, X., Boonstra, M. J., Cabrita, M., Perramon, J., Munive, S., Guala, A., Kincl, V., Haitjema, S., Dantas, C., & Asselbergs, F. W. (2024). Requirements for human-centered artificial intelligence: A heart failure study across Europe and Latin America. *Proceedings of the International Symposium on Medical Information Processing and Analysis, SIPAIM*. <https://doi.org/10.1109/SIPAIM62974.2024.10783548>

9. Journalism and online news consumption

A total of seven articles focused on journalism as a topic that were included in this analysis.

Two articles were located within the sample of 50 articles. After conducting a scan of the first 10 pages of Google Scholar and Perplexity.ai using relevant keywords from the initial search strategy, five more articles were located. A total of seven articles were included for the analysis.

There were 11 countries represented including Greece (n=1), Netherlands (n=2), Denmark (n=1), Romania (n=1), Poland (n=1), Switzerland (n=1), Germany (n=1), Belgium (n=1), Italy (n=1), Portugal (n=1), Spain (n=2). In terms of methodology, 1/7 was mixed-methods, 4/7 were qualitative, and 2/7 were quantitative studies. 3/7 articles had participants from more than one country. The total number of participants in the 7 articles was 2099.

For the purposes of this analysis, journalism is used as a proxy for online news consumption.

9.1 Overview

The growth of popularity of AI tools has the potential to transform the speed and connectedness of news (Simon et al., 2025; Sonni et al., 2024). However, there is also the concern that AI is eroding the necessity of traditional journalism practices which center reliability, truth, and trustworthy communication (Cools & Diakopoulos, 2024). Although AI may pose a direct threat to established journalistic processes, it is also seen as a necessity to utilize because of the potential time-saving and capacity-increasing benefits it can bring (Cools & Diakopoulos, 2024).

The overarching epistemic changes posed by AI-generated false news and deepfakes has applied pressure on journalistic values such as *integrity*, *truth* and *autonomy*. While it has traditionally been the role of the journalist to act as the gatekeeper to knowledge about the world and avoid disseminating biased or incorrect information, this has been affected by the introduction of AI-enabled tools which can rapidly generate text and media content (Guenther et

al., 2025). The risk of degradation of trust in journalism due to perceived AI use has made it pertinent for news companies to adapt to the current circumstances in ways that involve both distinguishing themselves as knowledge authorities and experimenting with AI tools to remain competitive in the news ecosystem (Sonni et al., 2024).

9.2 Normative values in journalism and AI

Various requirements posed by authors in the literature echo longstanding principles such as *truth, journalistic integrity, autonomy, and freedom of expression* (Cools & Diakopoulos, 2024; Cuartielles et al., 2023; Guenther et al., 2025; Kostarella et al., 2025; Mitova et al., 2025).

Values such as those in foundational documents like “Resolution 1003 on Ethical Principles for Journalism in Europe” (Parliamentary Assembly of the Council of Europe, 1993) appear to guide many of the expectations that journalists have for journalism in the context of AI use. Although many of the objectives in journalism have endured, the use of AI systems has shifted the way that these goals are realized (Figure 19).

Figure 19

Normative values in journalism from grey literature

Function	Values
What describes the ideal values embodied into news and opinions?	<p>Truth (para 5)¹ Verification and proof¹ Impartiality¹ Distinction between rumour and news¹ Accuracy</p> <p>Transparency (para 6)¹ Separate opinions and fact¹ Non-biasedness¹</p> <p>Professionalism (para 2)¹ Journalist rights and obligations¹</p> <p>Democracy (para 1)¹ Universalism²</p>

	<p>Right to information and communication¹ Ethical responsibility towards society¹</p> <p>Objectivity (para 5)¹ Neutrality¹ Impartiality¹ Distinction between truth and opinion¹</p>
<p>What describes values that characterize the right to information as a fundamental human right (for publishers, proprietors and journalists to align to)?</p>	<p>Lawfulness (para 8, para 10)^{3,1} Do not accept outside interference¹ Citizens' right to demand truthful information¹ Guard freedom of media¹ Separation between corporate structures and right to information¹</p> <p>Freedom of information (para 9, para 15)¹ Non-ownership of information¹ Information pluralism¹ Non-censorship¹ Freedom of expression¹ Do not exploit news for profit¹</p> <p>Transparency (para 12)¹ Accountability¹ Disclose identity of proprietors¹ Disclose economic interest in media¹</p> <p>Co-operation (para 13, para 14)¹ Collaboration between members of news ecosystem¹ Confidentiality of sources¹ Commitment to truth over ideological orientations¹ Truthful reporting¹ Ethical opinions¹ Respect fundamental right to information of citizens¹</p> <p>Freedom (para 14)¹ Freedom of expression of journalists¹</p> <p>Value of the individual (para 16)¹ Inherent human value of audience members and sources as individuals¹ Do not treat people as means⁴</p>

<p>What describes the role of ethical journalism in society?</p>	<p>Promoting democracy (para 17,18)¹ Promote civic engagement in public affairs¹ Provide information on public affairs (to public)¹ Contribute to role of information on public opinion¹ Civic freedoms¹</p> <p>Trustworthy communication (para 21)¹ Deliver honest opinions¹ Deliver truthful information¹ Do not exploit information for media purposes¹ Avoid conflicts of interest¹</p> <p>Respect and support for other public institutions (para 19, 20)¹ Do not replace role of public opinion¹ Do not unfairly shape public opinion¹ Do not misrepresent role of journalism¹ Avoid creating a “mediocracy”/Avoid transforming media and journalism into authorities¹</p>
<p>What describes the ideal behaviour of a journalist?</p>	<p>Privacy (para 23)¹ Respect privacy of individuals holding public posts¹ Proportionality (balance right to private life vs. freedom of expression)¹</p> <p>Fairness (para 22,25)¹ Respect presumption of innocence¹ Respect presumption of innocence in sub judge cases¹ Refrain from making judgments¹</p> <p>Lawfulness (para 25)¹ Obtain information by legal and ethical means</p> <p>Timely corrections (para 26)¹ Correct erroneous or false information in a timely manner¹</p> <p>Impartiality and independence (para 29, 30)¹ Ensure public and corporate relations which do not affect independence and impartiality of journalism¹</p>

	<p>Do not prioritize sensationalism or controversy over important news¹ Do not exploit duties for principal purpose of acquiring prestige or personal influence¹</p>
<p>Rules governing the media sector (companies and editorial staff)</p>	<p>Fair treatment of journalists in the workplace (para 28)¹ Fair working conditions¹ Fair pay¹</p> <p>Appropriate training (para 31)¹ Digital literacy¹ Timely and updated training¹</p> <p>Professionalism (para 32)¹ Co-operation between businesses, publishers, proprietors and journalists¹ Existence of editorial boards¹ Existence of professional rules for editorial staff¹</p> <p>Self-regulatory mechanisms (para 37) Multi-stakeholder involvement of guideline setting Respect mutually agreed upon ethical precepts Make ethical resolutions public</p>
<p>Rules governing situations of conflict and cases of special protection</p>	<p>Defend democratic values (para 33, 34)¹ Respect for human dignity¹ Solving problems by peaceful and tolerant means¹ Oppose violence¹ Avoid language that is hateful or confrontational¹ Reject all unfair discrimination¹ Avoid neutrality that comes at cost of democratic values¹ Encourage mutual understanding, tolerance and trust between communities at conflict¹</p> <p>Respect for children (para 35)¹ Avoid news that glorifies violence¹ Avoid exploiting sex¹ Avoid harmful consumerism¹ Avoid deliberately unsuitable language¹ Non-malevolence⁴</p> <p>Guard civic freedoms (para 36)¹ Guard freedom of expression of citizens¹</p>

	Guard fundamental right of citizens to receive truthful information and honest opinions ¹
--	--

Note. Based on a thematic extraction of:

- (1) Parliamentary Assembly of the Council of Europe. (1993, July 1). *Resolution 1003 (1993) on the ethics of journalism*. Council of Europe. <https://pace.coe.int/en/files/16414/html>
- (2) Schwartz, S. H. (2012). An Overview of the Schwartz Theory of Basic Values. *Online Readings in Psychology and Culture*, 2(1). <https://doi.org/10.9707/2307-0919.1116>
- (3) High-Level Expert Group on Artificial Intelligence. (2019). *Ethics Guidelines for Trustworthy AI*. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

Figure 20

Normative values extracted from empirical literature

Normative values extracted from articles	Sub-values/instrumental goals
Human oversight	Human expertise
	Human backed fact-checking
	Fight disinformation
	Domain-specific expertise
	Selection of news and topics
	Quality control
	Human verification
	Human control over hiring
Societal progress	Advancement of knowledge
	Augmentation of human skills
Role of AI in journalism	AI as a tool and not a replacement
	AI as a support
	AI as a creative partner
Transformation of journalism	Competitive edge
	Allow humans to do more
	International scalability
Democratization of knowledge	Multilingual translation
	Accessibility to information
Importance of human touch	Add context where needed
	Exercise human judgment
	Maintain personal relationship with readership/viewership
The enduring role of journalism in society	Adaptability as journalists
	AI literacy
	Uphold journalistic standards

Transparency	Disclosure of AI use
	Open disclosure about production and distribution practices
	Transparent practices and guidelines
Critical thinking	Identification of important world issues
	Supervision of AI work
Creativity	Producing original and meaningful content
	Human AI collaboration
Responsible AI	Guardrails
	AI policies and guidelines
	Journalist engagement in discovering and setting best practices
Proportionality	Balance pros and cons of AI-enabled tools
AI governance	Ethical and practical guidelines
	Stakeholder involvement and understanding of editorial needs
Reputation/integrity	Reputation of news outlet
	Reputation as a journalist
	Reputations of journalism
	Maintain journalistic values
Neutrality	Impartiality
	Prevent risk of bias in reporting
	Truthfulness
Accountability	Media company accountability
	Involvement of AI task force
	Evaluation of content
	Accountability of human opinion
	Accountability of selection of topics
	Accountable person
Reliability	Document critical accidents
	Rigor
	Supervised AI use
	AI as tool to double-check work

Values identified in the empirical literature about journalism and AI in the EU mirror many Trustworthy AI principles such as *accountability* and *human oversight* (2019) (Figure 26). Importance is placed on the responsibility of the journalist exercise human oversight on the use of AI in news as well as *societal progress*, *human touch*, *governance* and the *reputation* of journalists, organizations and journalism as a profession. These connect to the duty-based requirements outlined in Resolution 1003 (1993) which position the journalist as the gatekeeper and facilitator of information in society. The truthful dissemination of knowledge and its role in

supporting public institutions and structures in society is of paramount importance and appears to be linked to many instrumental goals such as *fight disinformation* and *quality control*.

However, the addition of values related to the human aspect of journalism are a reminder of a new goal, maintaining the survival and stability of the profession.

9.3 Analysis of main themes

9.3.1 Trustworthiness and the perception of truth

Recent advances in AI systems have shifted what it means to trust information presented online. AI-generated content such as videos, photos, audio or text has become increasingly realistic in recent years, interrupting traditional ways of separating fact from fiction. Bad actors may also use these tools to generate disinformation that supports false narratives. Even journalists that use AI tools in their own work risk inadvertently creating content that can contain hallucinations, unfair biases and falsehoods (Cuartielles et al., 2023).

Fact-checking assistance

On the other hand, AI systems can contribute to trust in news systems by being used to combat misinformation and disinformation. Fact-checkers state that LLMs like ChatGPT-3.5 can be a valuable resource for reporting, detecting falsehoods and debunking (Cuartielles et al., 2023). With human supervision, LLMs can provide assistance with fact-checking by being a soundboard for clues, initial starting points and ideas to begin the process (Cuartielles et al., 2023). AI systems can also help fact-checkers to save time by comparing information in real-time under time constraints (Cools & Diakopoulos, 2024). When used in conjunction with appropriate digital literacy and knowledge of the potential risks, AI systems can assist journalists in verifying information in high pressure situations.

Biases

However, journalists warn that AI systems do not always deliver accurate and reliable information. Many commercially available AI systems are limited to accessing information from a specific time period, meaning that information outputs are temporally limited (Cuartielles et al., 2023). In addition, AI systems may perform worse for particular geographic regions meaning that they are less useful for specific local or regional contexts (Cuartielles et al., 2023). Using AI-generated content may even create work for journalists who need to check the verity of the information before it is used (Guenther et al., 2025). Research has found that journalists are particularly wary of using AI tools for content generation without supervision (Cools & Diakopoulos, 2024; Drula, 2025). For these reasons, many journalists may be wary of trusting AI tools even if they could be of utility in some areas of work (Kostarella et al., 2025).

“I believe AI can assist with routine tasks, but I wouldn’t rely on such tools for synthesizing information at this stage, without further verification by an experienced journalist. (P2)” (Drula, 2025, p. 13)

Societal trust

There is the additional concern that the existence of fake news could degrade societal trust in institutions, negatively impacting population trust in traditional news sources and drive polarization (Cools & Diakopoulos, 2024). News publishers may also risk reputational damage if they publish AI-assisted/generated content that is then found to contain hallucinations or falsehoods (Cools & Diakopoulos, 2024). However, disapproval of AI use in general can affect readers’ trust of news publishers that use AI-enabled tools (Cools & Diakopoulos, 2024; Kostarella et al., 2025). Awareness of the possibility of AI-generated misinformation may thus affect audience trust of news in general, with or without evidence of false information (Cools & Diakopoulos, 2024).

D3.1 Impact Assessment of Values Affected

A multi-pronged approach involving the promotion of AI literacy and responsible AI use can be important to mitigating negative impacts on trust in news. Additionally, encouraging audience members to seek out verified news sources rather than turning to alternative sources due to misplaced distrust in news sources or misinformation is pertinent (Kostarella et al., 2025).

9.3.2 Impacts on underrepresented groups

Risk of disinformation

AI use in news has the potential to unintentionally misrepresent truth and amplify falsehoods to the detriment of vulnerable groups (Cools & Diakopoulos, 2024). Content generated with the assistance of AI may contain embedded unfair biases or hallucinations that contribute to existing biases and stereotypes and propagate false narratives (Cools & Diakopoulos, 2024). Furthermore, disinformation is able to be created and disseminated more readily with AI systems like LLMs that can translate and distribute messages rapidly (Cuartielles et al., 2023). Marginalized groups as well as ethnic and gender minorities could therefore be disproportionately targeted by both inherent biases in AI content as well as the creation of disinformation.

Democratization of knowledge

Studies mention that AI can also be used in ways to democratize knowledge and information by facilitating the distribution and creation of news content and increasing the capacity of journalism professionals. This is especially salient to the case of science journalism where journalists can use AI systems to synthesize complex scientific knowledge into accessible content for the layperson (Guenther et al., 2025). AI can also be used for translation purposes, facilitating the distribution of news internationally (Guenther et al., 2025). This may assist underrepresented groups and regions with communicating news and expanding their reach (Guenther et al., 2025).

D3.1 Impact Assessment of Values Affected

AI was also seen as making research from different parts of the world, such as Africa, more visible, or research from smaller universities, and this could be linked to access and analysis of large data sets ... linked to increasing open access practices. A wish shared by many was that through using AI and genAI, science news could become more diverse, more in-depth, and more tailored to the needs of the audience. (Guenther et al., 2025, p. 7)

9.3.3 The changing nature of journalism

The growing capabilities of AI-enabled tools to produce audio and visual media and written content has shifted the way that journalists produce and curate content, leaving room for experimentation, expansion of the role of journalists as well as guideline setting within newsrooms (Kostarella et al., 2025; Mitova et al., 2025). The type of connections that journalists can form with their audiences is also changing with possibilities to personalize news and deliver content in novel ways (Cools & Diakopoulos, 2024). This may present an opportunity for journalism to define itself in new ways, along with risks to traditional journalistic practices.

Journalists' voices

For some, the rise of AI is seen as an opportunity to re-emphasize the value of journalists' creative and authentic voice and set themselves apart from AI-generated content (Cools & Diakopoulos, 2024). There is the possibility that propagation of AI-generated content in the news sphere could kindle a desire for traditional forms of journalism that have less AI involvement (Cools & Diakopoulos, 2024). From this perspective, the introduction of AI-enabled tools represents less of a threat to the field of journalism than other narratives of job loss, outsourcing and replacement (Cools & Diakopoulos, 2024).

Connection with the audience

Journalists in studies emphasize that aspects of journalism related to human connection and creative style are both personally important to their identity as well as to readers in the context

of AI use in news (Cuartielles et al., 2023; Guenther et al., 2025). These concerns are also tied closely to journalistic autonomy and how journalists envision their role in society (Cools & Diakopoulos, 2024). Some see the possibility of AI use in journalism as a potential threat to the ability for them to connect to their audience as well as other news professionals and sources (Guenther et al., 2025). The use of AI which emulates one's "voice" and writing style may be both useful and an uncomfortable possibility (Guenther et al., 2025). Journalists also doubt the ability of AI to fully understand the human experience which is posited as a requisite to quality journalism for some (Guenther et al., 2025).

Labour impacts

Some aspects of journalism are less likely to be replaced by automation such as those involving human oversight like fact-checking and selection of topics and ideas (Cuartielles et al., 2023; Guenther et al., 2025). However, journalists have also expressed that there is the necessity to adapt to the changing landscape of news consumption in order to avoid becoming obsolete (Cuartielles et al., 2023; Drula, 2025; Guenther et al., 2025; Kostarella et al., 2025). In particular, journalists in low-resourced media companies as well as those unwilling or unable to adapt to changes in the industry can be at risk of losing their jobs (Guenther et al., 2025).

Navigating the role that journalists play in society with the effects of AI on the consumption and distribution of information may require trial and error. Balancing experimentation with the potential harms that result from using AI will also be challenge for journalists as they make decisions about best practices in their field.

9.3.4 Accountability and standard setting

Furthermore, there is the question about how to ascribe accountability to regarding the spectrum of AI-assisted to AI-generated articles. However, for the most part, the studies in this sample describe journalists' desire to use AI as a tool or support, not as a replacement for

generating original content (Cools & Diakopoulos, 2024; Cuartielles et al., 2023; Drula, 2025; Guenther et al., 2025; Mitova et al., 2025). The theme of authorship appears less often in the literature than the responsibility of journalists to ensure that the tools they are using are not interfering with the communication of truthful and unbiased information.

Overreliance

Many sentiments in the articles relate to concerns about the impact of AI-enabled tools on the quality of news. It is for this reason that some journalists are hesitant about relying on AI excessively to assist them with their tasks, at the risk of producing work which does not align with their journalistic values or expectations for quality (Cools & Diakopoulos, 2024). Other risks of AI use such as hallucinations, lack of sources, superficial content and misinformation inherently affect the desired quality of news. Thus, there seems to be little support of handing off journalistic responsibility when using AI-enabled tools but rather, a strong push for human oversight and integrity (Guenther et al., 2025).

Inconsistent regulations and guidelines

However, the literature shows that journalists feel that inconsistent regulation and guidelines on AI disclosure and use, as well as implementation of AI-enabled tools pose a more immediate risk to accountability (Dijkstra et al., 2024; Drula, 2025; Guenther et al., 2025). The current shortage of best practices and harmonized standards means that journalists and news companies may be missing guidance on how to proceed next with AI tools (Cools & Diakopoulos, 2024). Responsible AI requires the consideration of editorial staff needs and setting ethical guidelines that are applicable for the daily work of staff (Drula, 2025). As reported in Resolution 1003 (par 37), it is imperative for news ecosystems to engage in mutually agreed upon reporting and accountability measures for journalism that are also public (1997).

9.3.5 Impacts on local/small/regional news

Financial challenges

The literature shows that small news publishers such as local and regional news seem to experience different threats and impacts than larger news outlets from AI. Journalists mention that small media companies that are financially struggling may put greater energy into saving costs by outsourcing work to AI (Guenther et al., 2025). For example, collapse of small media companies in previous economic recessions put regions in the EU such as small communities in Greece in “news deserts” where regions were left without local news (Kostarella et al., 2025). Furthermore, local companies may also have less capital to invest into AI experimentation, which has been described as a priority in the changing news climates (Kostarella et al., 2025). Additionally, many local news publishers in the EU may also have to cater to language and cultural needs that mainstream LLMs and AI systems do not perform as well for. Altogether, these factors could put small and regional news companies at a disadvantage against larger media organizations that have more leeway to find out how to optimize AI use for their needs.

Improving workflow and efficiency

AI systems can also assist with carrying out with tasks such as translation, transcription and news distribution, potentially assisting smaller companies with news dissemination and personalization (Cools & Diakopoulos, 2024; Cuartielles et al., 2023; Guenther et al., 2025; Kostarella et al., 2025). As well, AI can also be used to scale work processes and content creation, potentially increasing the capacity of smaller news companies (Cuartielles et al., 2023). The use of AI tools can help level the playing field in smaller media companies for those that have staff with skillsets to utilize AI.

AI literacy and digital divides

Some research has shown that small news companies are often not aware of how to use AI tools, with usage and interest widely stratified by age, education and journalist role (Kostarella et al., 2025). Efforts put into developing AI literacy for older adult journalists or journalists without access to useful tools can help bridge the gap and mitigate any risks to their roles.

9.3.6 Journalistic autonomy and involvement in AI development

The importance of including journalists and other news professionals into the process of developing AI systems and guidelines is echoed in both the literature (Cuartielles et al., 2023) as well as earlier soft policy (Parliamentary Assembly of the Council of Europe, 1993). However, some journalists have stated that commercially available systems and developers have not consulted professions such as fact-checkers in the development process (Cuartielles et al., 2023). Involving the input of journalists may improve the effectiveness of AI-enabled tools as well as leverage their expertise in preventing harms such as disinformation and fake news from occurring.

The importance of utility

The usefulness of AI-enabled tools in optimizing workflow, assisting with repetitive tasks and making information more accessible in different languages and formats may be of value to journalists who feel under pressure to remain productive and competitive in a changing media landscape (Guenther et al., 2025; Kostarella et al., 2025). Most of the trade-offs are associated with the utility of AI systems and the risks of using AI systems in a media context such as unreliability, lack of transparency and the risk of propagating misinformation (Figure 28). There is also the possibility that the usefulness of AI-enabled tools may reduce the demand for journalists and threaten the careers of journalists. While values derived from Resolution 1003

are mostly related to duty, the relevant trade-offs identified in the literature seem to affect utilitarian concerns about the nature of work.

Figure 21

Opportunities for use of AI in journalism in the empirical literature

Instrumental uses for AI in journalism	Sub-uses
Fact-checking	Debunking
	Detection of falsehoods
	Real-time fact checking
Using AI in the workplace	Efficiency
	Assist with workflow
	Save time
	Reduce menial and repetitive tasks
	Allow focus on meaningful and original content
	Scalability
	Assist with news writing
	Analysis of data
	Assist with producing content
	Assist with translation
	Assist with content distribution
	Assist with administrative tasks
	Assist with editing
	Finding stories
Summarize information	
Search engine	
Utility for audience	Interactivity
	Increased accessibility of news
	Content personalization
	User engagement

Figure 22

Vulnerable populations in journalism

Vulnerable population	Explanation
------------------------------	--------------------

Small and regional journalism	Less financial resources to cope with changes in journalism
Marginalized, minority and vulnerable groups	Embedded risks of biases and stereotypes in training datasets Polarization of opinions Targeted disinformation
Older journalists	Less familiar with digital tools Less likely to accept AI use More unwilling to adapt
Young journalists	Risk of outsourcing entry-level tasks

9.4 References for journalism/online news consumption use-case

- Cools, H., & Diakopoulos, N. (2024). Uses of generative AI in the newsroom: Mapping journalists' perceptions of perils and possibilities. *Journalism Practice*, 0(0), 1–19.
<https://doi.org/10.1080/17512786.2024.2394558>
- Cuartielles, R., Ramon-Vegas, X., & Pont-Sorribes, C. (2023). Retraining fact-checkers: The emergence of ChatGPT in information verification. *Profesional de la Informacion*, 32(5).
<https://doi.org/10.3145/epi.2023.sep.15>
- Dijkstra, A. M., de Jong, A., & Boscolo, M. (2024). Quality of science journalism in the age of Artificial Intelligence explored with a mixed methodology. *PLOS ONE*, 19(6).
<https://doi.org/10.1371/journal.pone.0303367>
- Drula, G. (2025). The human-AI partnership in Romanian newsrooms: AI as both a news topic and a tool. *Journalism Practice*, 1–20. <https://doi.org/10.1080/17512786.2025.2513429>
- Guenther, L., Kunert, J., & Goodwin, B. (2025). “Away from this duty of chronicler and towards the unicorn”: How German science journalists assess their future with (generative) Artificial Intelligence. *Journal of Science Communication*, 24(2), A06.
<https://doi.org/10.22323/2.24020206>
- Kostarella, I., Saridou, T., Dimoulas, C., & Veglis, A. (2025). Can Artificial Intelligence (AI) Spring an Oasis to the Local News Deserts? *Journalism Practice*, 19(10), 2341–2361.
<https://doi.org/10.1080/17512786.2025.2513423>
- Mitova, E., Blassnig, S., Strikovic, E., Urman, A., Hannak, A., Vreese, C. de, & Esser, F. (2025). Exploring Public Attitudes Toward Generative AI for News Across Four Countries.



HORIZON-MSCA-2023-DN-01
Grant Number 101169473



Journal of Quantitative Description: Digital Media, 5.

<https://doi.org/10.51685/jqd.2025.012>

Section 3: Overview of findings

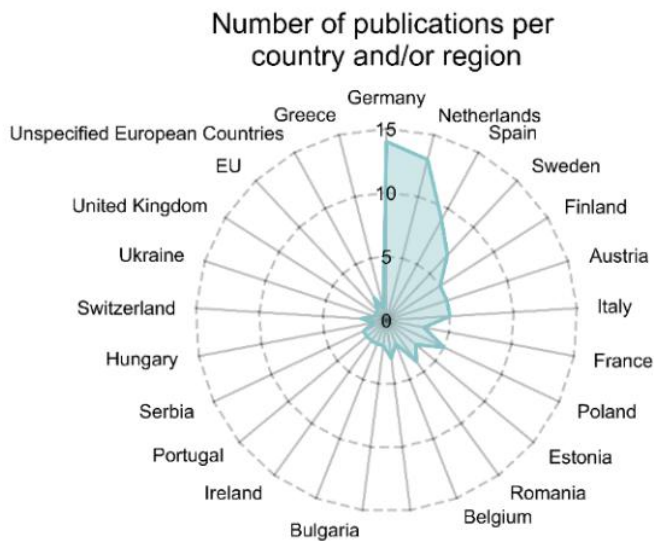
10. Bibliometric analysis of empirical literature

10.1 Result and discussion of bibliometric analysis results

10.1.1 Geographic bias towards Western Europe

Figure 23

Number of publications per country and/or region



Note. Studies may be represented multiple times if they contain multiple populations.

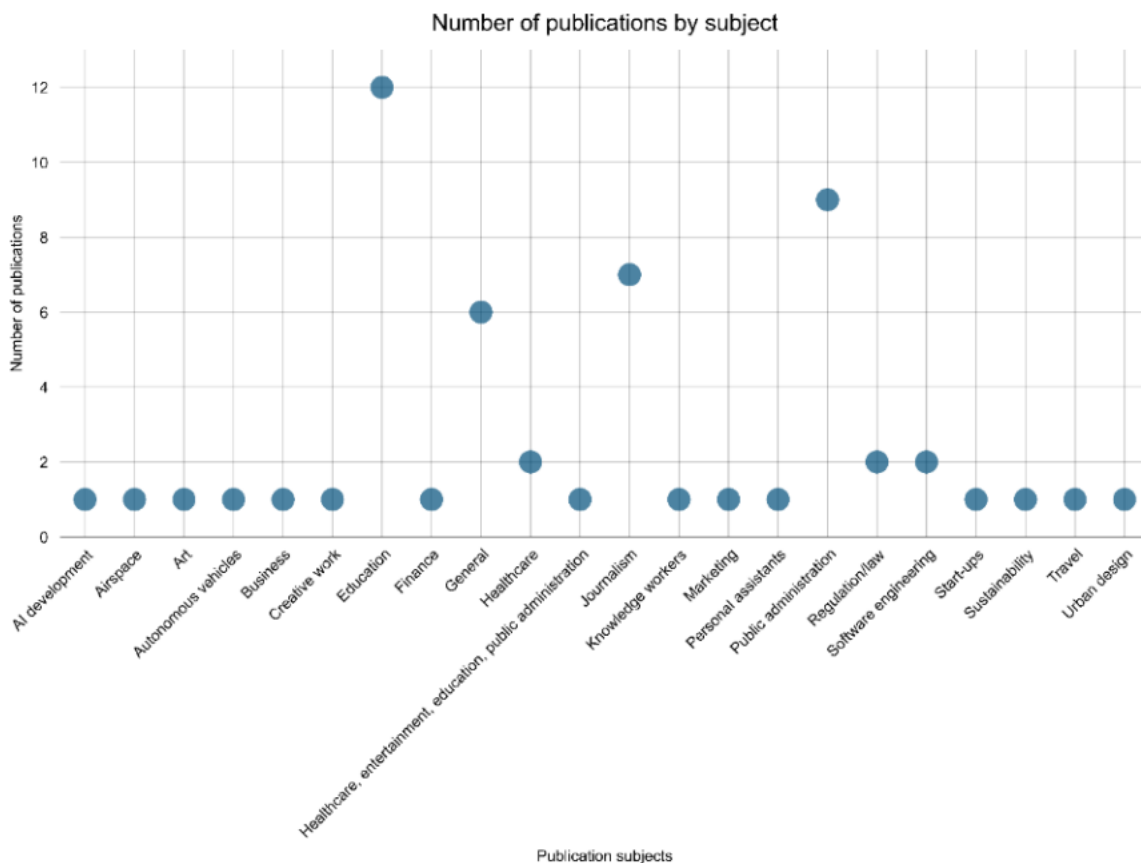
In the sample of 55 studies total, a disproportionate number of articles were published in Germany (n=15), Netherlands (n=13) and Spain (n=9) (Figure 23). In decreasing order, number of articles published in: Sweden (n=7), Finland (n=5), Austria (n=5), Italy (n=5), Poland (n=5), Romania (n=4), France (n=3), Denmark (n=3), Estonia (n=3), Bulgaria (n=2), Belgium (n=2), Czech Republic (n=2), Ireland (n=2), Portugal (n=2), Serbia (n=2), Hungary (n=1), Ukraine (n=1), Greece (n=1). Some studies included countries outside of the EU alongside countries in the EU; Switzerland (n=2), United Kingdom (n=1), EU (general/unspecified country) (n=1) and Europe (general/unspecified country) (n=2).

The overrepresentation of these countries in the preliminary sample of 55 papers correspond to previous studies on AI ethics guidelines which indicate that their samples contained a larger proportion of documents from Western European countries like Germany (Corrêa et al., 2023; Fjeld et al., 2020). Future studies could strive to include values from Eastern EU countries in their investigation for a more diverse and representative analysis of EU societal values.

10.1.2 Diversity of subject areas

Figure 24

Number of publications by subject



Note. Studies may be represented multiple times if they contain multiple subjects

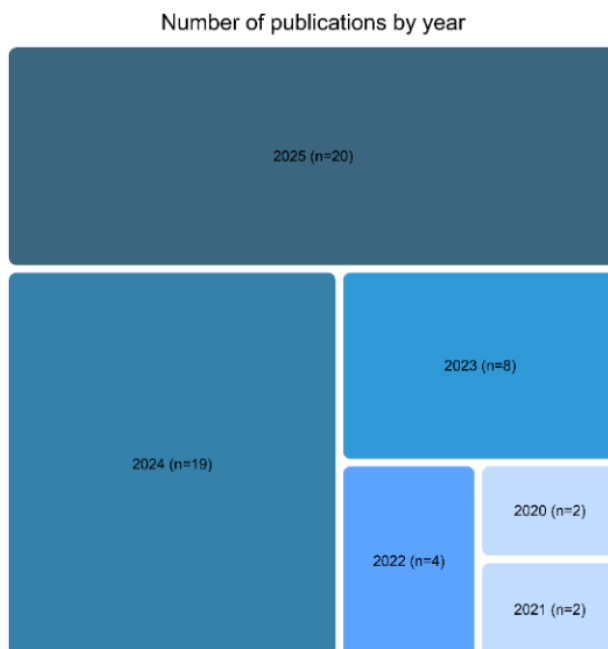
A large portion of studies was found in the realms of education (n=12) and on general population perspectives on AI values without subject focus (n=6) (Figure 24). Two studies were found each in the fields of healthcare, regulation and software engineering. There were two studies found in the field of journalism, but an additional Google scholar search revealed five more studies bringing the total up to seven. An added Google scholar search in healthcare revealed no additional relevant studies.

As the search for articles was conducted without keywords from specific subject areas, it is possible that a more targeted search towards specific journals and sectors may reveal more articles. Additionally, the searches were conducted between September 2025 and November 2025, meaning that additional articles could have been published between the end of the search period and December 2025.

10.1.3 Publications by year

Figure 25

Number of publications by year

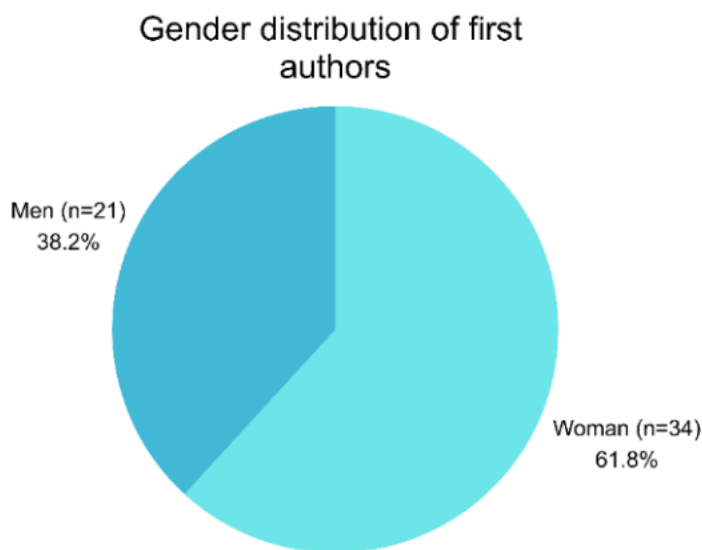


Although the search spanned 2017-2025, only studies after 2020 were represented in the preliminary sample of 55 full texts with the majority of studies being published in 2025 (n=2) and 2024 (n=19) (Figure 25). This speaks to a growing body of empirical literature exploring values in EU populations and potential interest in value-alignment efforts.

10.1.4 Representation of women in authorship

Figure 26

Distribution of gender of first authors



The gender distribution of first authors showed a slightly higher percentage of women (n=34) in comparison to men (n=21) (Figure 26). This runs against trends previously observed in first authorship of papers in the AI community (Ding et al., 2025; Yuan et al., 2024) as well as the philosophy community (Hassoun et al., 2022).

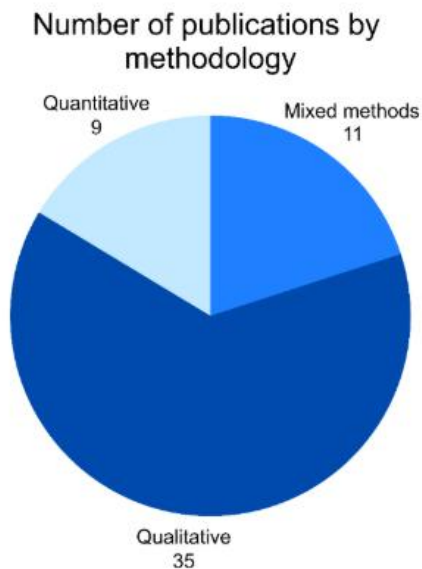
As well, findings regarding the percentage of first authors who were women vs. men show approximately 23% more first authors who were women vs men in the sample of 55 papers (Figure 10). This is counter to previously observed trends in academic publications in the AI

community as well as the philosophy community where first authors who are men appeared disproportionately more (Ding et al., 2025; Hassoun et al., 2022; Yuan et al., 2024). While these results may represent only a portion of growing publications in the field (see Figure 25 for the number of publications per year in the sample), this trend may deserve exploration into the gender representation of authors in AI ethics literature.

10.1.5 Qualitative focus

Figure 27

Number of publications by methodology



Additionally, the majority of the studies were qualitative ($n=35$) with only nine quantitative and eleven mixed methods (Figure 27). This may be due to the nature of the research question which explores values, perspectives and principles, which may be more readily explored via qualitative methods such as open-ended surveys and interviews.

The findings included were also solely extracted from insights from EU populations and empirical results in articles. Results from quantitative explorations such as surveys and

questionnaires were excluded if EU populations were aggregated with other non-EU, non-UK and non-CH populations.

As qualitative results may be more easily separated by participant insight in many studies, it may be the case that qualitative studies were overrepresented due to the method of search and extraction in this analysis. On the other hand, future research that requires quantitative results about EU societal values could face challenges if relying on academic literature. Grey literature such as Eurobarometer surveys (European Commission, 2024) and other reports from private and public organizations (Google 2025; UNESCO, 2025) may provide more quantitative results.

11. Key takeaways and next steps

This work intends to build on a body of previous literature which highlights that value-alignment efforts necessitate the involvement of both societal and technical methods in order to further societal goals (Amugongo et al., 2023; Ferretti, 2024; Hagendorff, 2020; Shen et al., 2024). The definitions provided by these textual analyses sought to combine normative expectations in different areas of AI ethics with empirical literature to summarize relevant values into sub-dimensions and actionable recommendations. These values inform a normative framework for societal value alignment that articulate how core human values and needs such as those mentioned in section 3.2 (Rokeach, 1973; Schwartz, 2012) are transformed into AI-specific motivators and needs during technical value alignment and societal processes.

Furthermore, this research has shed light on the emerging body of work that has defined some relevant human values in AI ethics (see section 1). While these values provide a direction for implementation, they need to be combined with best practices and preferences that define how normative principles manifest for use-cases. It is necessary to maintain a connection between abstract values and real-life practices to continually and iteratively define shared human desires and needs. The hope is that in viewing value-alignment as an inherently sociotechnical process

that is a mediator between individual and societal goals, practitioners can parse out the differences between values and apply them correctly during implementation and practice.

11.1 Future directions

The preliminary values identified summarize relevant empirical and grey literature about EU societal values and AI. Future work can explore the values and definitions provided in this document to organize them further into coherent categories for application in policy and technical requirements.

From a policy perspective, categories of values identified such as those in Figure 8 can provide the basis of future work that examines how values can be actionable in AI governance. For instance, values like *lawfulness* can be fostered through non-technical methods such as encouraging alignment with AI ethical guidelines in soft policy and examining the impacts of regulations like the EU AI Act on rights and freedoms. Furthermore, incorporating normative values like *trustworthiness* into context-specific guidelines, such as in sectors like education and healthcare can aid in providing guidance that works on the ground-level. In doing so, identified sub-goals could assist in separating different dimensions of values that can then be used to apply policy and ethical guidelines.

From a technical perspective, these lists of values can be used to give a general idea of which EU societal values are salient for value-alignment efforts, as well as a closer glimpse into how specific values may be articulated in use-cases. One future direction is to use selected value-sets and defined preferences as “ground truths” to shape an evidence-informed framework that can be used for the fine-tuning of context-specific, EU societal-value aligned LLMs. Values and sub-values can also be utilized for strategies such as building value-alignment system prompts that embody value sets, building preference pairs that correspond values and their dimensions, or generating rules that reflect established ethical frameworks in specific use-cases and

professions. Next steps can involve a more detailed exploration into how values interact with each other and which ones can be prioritized within value sets.

As well, technical specifications can be combined with explorative, participatory methods which combine human preferences identification with value-defining exercises. For instance, engaging with stakeholders in different use cases can offer insight that helps contextualize what values mean to people on the ground. Other methods such as participatory workshops or surveys can be used to gather information about how to translate abstract principles into implementation.

Furthermore, trends in the bibliometric analysis of 55 empirical studies from the preliminary systematic review results show interesting discrepancies in the types of publications found with findings about EU societal values in AI. Next steps include exploring representation of a more diverse set of EU countries in regard to values, particularly Eastern European countries', further research examining and mapping the proportion of first authors who are women compared to men in the literature on EU societal values of AI and quantitative work to build on findings from exploratory qualitative studies.

11.2 Conclusions

This paper provides a preliminary and broad overview of values regarding AI as they appear in empirical studies and normative rules and guidelines that apply to the EU context. Drawing from collaborations between DCs in the alignAI project to build a repository of normative guidelines and preliminary evidence from an ongoing systematic review, studies and grey literature were coded for relevant human values, sub-values, risks, opportunities, trade-offs and vulnerable populations. The findings show that there is a great deal of conceptual overlaps between iterations of human values important in general AI ethics guidelines as well as specific use cases in education, healthcare and journalism.



Furthermore, a straightforward, replicable methodology for value identification based on Rokeach and Schwartz's value theories across different disciplines and documents was used to separate broad, abstract values from more instrumental goals and sub-values. Doing so allowed values to be further split into other dimensions such as risks, opportunities, and trade-offs. Next steps will involve further examination of values and trade-offs to further value alignment efforts in LLMs developed for the three-use cases.

11.3 Disclosure of AI use

Claude Sonnet 4.6, Gemini 3 and Chat GPT-5 were used to preliminarily organize data into tables, proof-reading for grammar and phrasing, and checking references. All AI generated content was double-checked and the author takes full responsibility for the content written in this document.

References

- Adžić, S., Savic-Tot, T., Vukovic, V., Radanov, P., & Avakumović, J. (2024). Understanding student attitudes toward GenAI tools: A comparative study of Serbia and Austria. *International Journal of Cognitive Research in Science, Engineering and Education*, 12(3), 123–145. <https://doi.org/10.23947/2334-8496-2024-12-3-583-611>
- American Psychological Association. (2017). *Ethical principles of psychologists and code of conduct (2002, amended effective June 1, 2010, and January 1, 2017)*. <https://www.apa.org/ethics/code/>
- Amugongo, L. M., Bidwell, N. J., & Corrigan, C. C. (2023). Invigorating Ubuntu Ethics in AI for healthcare: Enabling equitable care. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, 583–592. <https://doi.org/10.1145/3593013.3594024>
- Balcioğlu, Y. S., Çelik, A. A., & Altındağ, E. (2025). A turning point in AI: Europe's human-centric approach to technology regulation. *Journal of Responsible Technology*, 23, Article 100128. <https://doi.org/10.1016/j.jrt.2025.100128>
- Beauchamp, T. L., & Childress, J. F. (1979). *Principles of biomedical ethics*. Oxford University Press.
- Billah, M. M., Hamjaya, H. S., Shiralizade, H., Singh, V., & Inam, R. (2025). Large language models' trustworthiness in the light of the EU AI Act—A systematic mapping study. *Applied Sciences*, 15(14), Article 7640. <https://doi.org/10.3390/app15147640>
- Bucea-Manea-Țoniș, R., Kuleto, V., Gudei, S. C. D., Lianu, C., Lianu, C., Ilić, M. P., & Paun, D. (2022). Artificial intelligence potential in higher education institutions enhanced learning environment in Romania and Serbia. *Sustainability*, 14(10). <https://doi.org/10.3390/su14105842>
- Cernadas, E., & Fernández-Delgado, M. (2021). Embedded ethics to teach machine learning courses: An experience. *2021 XI International Conference on Virtual Campus (JICV)*, 1–4. <https://doi.org/10.1109/JICV53222.2021.9600426>
- Cools, H., & Diakopoulos, N. (2024). Uses of generative AI in the newsroom: Mapping journalists' perceptions of perils and possibilities. *Journalism Practice*, 0(0), 1–19. <https://doi.org/10.1080/17512786.2024.2394558>
- Corrêa, N. K., Galvão, C., Santos, J. W., Pino, C. D., Pinto, E. P., Barbosa, C., Massmann, D., Mambrini, R., Galvão, L., Terem, E., & Oliveira, N. de. (2023). Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, 4(10). <https://doi.org/10.1016/j.patter.2023.100857>

- Cuartielles, R., Ramon-Vegas, X., & Pont-Sorribes, C. (2023). Retraining fact-checkers: The emergence of ChatGPT in information verification. *Profesional de la Informacion*, 32(5). <https://doi.org/10.3145/epi.2023.sep.15>
- D'Alessandro, W. (2025). Deontology and safe artificial intelligence. *Philosophical Studies*, 182(7), 1681–1704. <https://doi.org/10.1007/s11098-024-02174-y>
- del Álamo Cienfuegos, A., Lis, P., Zadorozhna, O., Espiritusanto, O., Gawronska-Nowak, B., & Zarzycka, A. (2024). AI: What do we fear? What do we hope for? Perception of the societal impact of AI in a European transnational cross-sectional study. *2024 IEEE International Conference on Engineering, Technology, and Innovation (ICE/ITMC)*, 1–9. <https://doi.org/10.1109/ICE/ITMC61926.2024.10794321>
- Dijkstra, A. M., de Jong, A., & Boscolo, M. (2024). Quality of science journalism in the age of artificial intelligence explored with a mixed methodology. *PLOS ONE*, 19(6). <https://doi.org/10.1371/journal.pone.0303367>
- Ding, Y., Liu, J., Lyu, Z., Zhang, K., Schoelkopf, B., Jin, Z., & Mihalcea, R. (2025). *Voices of her: Analyzing gender differences in the AI publication world* (arXiv:2305.14597). arXiv. <https://doi.org/10.48550/arXiv.2305.14597>
- Dobre, C., & Dobre, C. R. (2021). The actuality of Aristotelian virtues. *Filosofya-Philosophy*, 30(3), 259–269.
- Drula, G. (2025). The human-AI partnership in Romanian newsrooms: AI as both a news topic and a tool. *Journalism Practice*, 1–20. <https://doi.org/10.1080/17512786.2025.2513429>
- European Commission. (2024). *Artificial intelligence and the future of work*. <https://europa.eu/eurobarometer/surveys/detail/3222>
- European Federation of Psychologists Associations. (2025). *European Federation of Psychologists Associations meta code of ethics*. https://www.efpa.eu/sites/default/files/2025-06/2025-06-03_efpa-meta-code-of-ethics.pdf
- Felkner, V. K., Chang, H.-C. H., Jang, E., & May, J. (2024). *WinoQueer: A community-in-the-loop benchmark for anti-LGBTQ+ bias in large language models* (arXiv:2306.15087). arXiv. <https://doi.org/10.48550/arXiv.2306.15087>
- Ferretti, T. (2024). Value Alignment Without Institutional Change Cannot Prevent the Societal Risks of Artificial Intelligence. *LSE Public Policy Review*, 3(3). <https://doi.org/10.31389/lseppr.113>
- Figueras, C., Rossitto, C., & Cerratto Pargman, T. (2024). Doing responsibilities with automated grading systems: An empirical multi-stakeholder exploration. *Proceedings of the 13th Nordic Conference on Human-Computer Interaction, NordiCHI '24*. <https://doi.org/10.1145/3679318.3685334>

- Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI* (SSRN Scholarly Paper No. 3518482). Social Science Research Network. <https://doi.org/10.2139/ssrn.3518482>
- Floridi, L., Cowsls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- Gabriel, I. (2020). Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3), 411–437. <https://doi.org/10.1007/s11023-020-09539-2>
- Gooding, P., van Kolfschooter, H., & Centola, F. (2024). Artificial Intelligence in mental healthcare. *Mental Health Europe*. <https://easped.eu/news-detail/ensuring-a-responsible-use-of-artificial-intelligence-in-mental-healthcare/>
- Google. (2025). *The future report: Perspectives on technology from teenagers in Europe* (p. 77). <https://futurereport.eu/>
- Gouveia, V. V., Milfont, T. L., & Guerra, V. M. (2014). The functional theory of human values: From intentional overlook to first acknowledgement—A reply to Schwartz (2014). *Personality and Individual Differences*, 68, 250–253. <https://doi.org/10.1016/j.paid.2014.03.025>
- Guenther, L., Kunert, J., & Goodwin, B. (2025). “Away from this duty of chronicler and towards the unicorn”: How German science journalists assess their future with (generative) artificial intelligence. *Journal of Science Communication*, 24(2), A06. <https://doi.org/10.22323/2.24020206>
- Guo, H., Yao, J., Zhou, X., Yi, X., & Xie, X. (2025). *Counterfactual reasoning for steerable pluralistic value alignment of large language models* (arXiv:2510.18526). arXiv. <https://doi.org/10.48550/arXiv.2510.18526>
- Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1), 99–120. <https://doi.org/10.1007/s11023-020-09517-8>
- Han, S., Kelly, E., Nikou, S., & Svee, E.-O. (2022). Aligning artificial intelligence with human values: Reflections from a phenomenological perspective. *AI & Society*, 37(4), 1383–1395. <https://doi.org/10.1007/s00146-021-01247-4>
- Hanel, P. H. P., Litzellachner, L. F., & Maio, G. R. (2018). An empirical comparison of human value models. *Frontiers in Psychology*, 9. <https://doi.org/10.3389/fpsyg.2018.01643>
- Hassoun, N., Conklin, S., Nekrasov, M., & West, J. (2022). The past 110 years: Historical data on the underrepresentation of women in philosophy journals. *Ethics*, 132(3). <https://www.journals.uchicago.edu/doi/10.1086/718075>

- High-Level Expert Group on Artificial Intelligence. (2019). Ethics Guidelines for Trustworthy AI. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- Jemetz, M., Dolezal, D., & Motschnig, R. (2024). Secondary teachers' self-perceived AI competences in relation to renowned European digital competence frameworks. *Informatics in Schools: Innovative Approaches to Computer Science Teaching and Learning, ISSEP 2024, Lecture Notes in Computer Science, 15228*, 3–17. https://doi.org/10.1007/978-3-031-73474-8_1
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Kant, I. (2018). *Groundwork for the metaphysics of morals: With an updated translation, introduction, and notes* (A. W. Wood, Trans. & Ed.). Yale University Press. (Original work published 1785)
- Kalving, M., Colley, A., & Häkkinen, J. (2024). Where AI and Design Meet—Designers' Perceptions of AI Tools. Proceedings of the 13th Nordic Conference on Human-Computer Interaction, NordiCHI '24. <https://dl.acm.org/doi/10.1145/3679318.3685389>
- Klemettilä, P. A., Sharma, S., Mochiyama, F., Iivari, N., Iwata, M., & Koivisto, J. (2025). “It’s just a machine that predicts”—Demystifying artificial intelligence/machine learning with teenagers. *Proceedings of the 24th Interaction Design and Children, IDC '25*, 168–182. <https://doi.org/10.1145/3713043.3728853>
- Kostarella, I., Saridou, T., Dimoulas, C., & Veglis, A. (2025). Can artificial intelligence (AI) spring an oasis to the local news deserts? *Journalism Practice, 19*(10), 2341–2361. <https://doi.org/10.1080/17512786.2025.2513423>
- Kraut, R. (2022). Aristotle's ethics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2022 ed.). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/fall2022/entries/aristotle-ethics/>
- Leoste, J., Pöial, J., Kivisalu, E., Marjanovic, U., Rakic, S., & Robal, T. (2025). Integration of artificial intelligence in higher education programming courses: Insights from student perspectives and practices. *Lecture Notes in Networks and Systems, 1260*, 417–427. https://doi.org/10.1007/978-3-031-85652-5_42
- Maris, M. T., Koçar, A., Willems, D. L., Pols, J., Tan, H. L., Lindinger, G. L., & Bak, M. A. R. (2024). Ethical use of artificial intelligence to prevent sudden cardiac death: An interview study of patient perspectives. *BMC Medical Ethics, 25*(1). <https://doi.org/10.1186/s12910-024-01042-y>
- Mitova, E., Blassnig, S., Strikovic, E., Urman, A., Hannak, A., Vreese, C. de, & Esser, F. (2025). Exploring public attitudes toward generative AI for news across four countries. *Journal of Quantitative Description: Digital Media, 5*. <https://doi.org/10.51685/jqd.2025.012>

- Miao, F., Holmes, W., Ronghuai, H., & Zhang, H. (2021). AI and education: Guidance for policy-makers. <https://unesdoc.unesco.org/ark:/48223/pf0000376709>
- OECD. (2021). State of implementation of the OECD AI Principles: Insights from national AI policies (No. 311; OECD Digital Economy Papers). Organisation for Economic Co-operation and Development (OECD). <https://doi.org/10.1787/1cd40c44-en>
- Parliamentary Assembly of the Council of Europe. (1993, July 1). *Resolution 1003 (1993) on the ethics of journalism*. Council of Europe. <https://pace.coe.int/en/files/16414/html>
- Pavlova, Y., & Slavov, V. (2025). Exploring the ethical use of AI technologies/applications in academic environments (M. Kostov & M. Atanasovski, Eds.). *2025 60th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)*, 1–4. <https://doi.org/10.1109/ICEST66328.2025.11098253>
- Pham, B.-C., & Davies, S. R. (2025). What problems is the AI act solving? Technological solutionism, fundamental rights, and trustworthiness in European AI policy. *Critical Policy Studies*, *19*(2), 318–336. <https://doi.org/10.1080/19460171.2024.2373786>
- Prather, J., Denny, P., Leinonen, J., Becker, B. A., Albluwi, I., Craig, M., Keuning, H., Kiesler, N., Kohn, T., Luxton-Reilly, A., MacNeil, S., Petersen, A., Pettit, R., Reeves, B. N., & Savelka, J. (2023). The robots are here: Navigating the generative AI revolution in computing education. *Proceedings of the 2023 Working Group Reports on Innovation and Technology in Computer Science Education, ITiCSE-WGR '23*, 108–159. <https://doi.org/10.1145/3623762.3633499>
- Puig-Bosch, X., Boonstra, M. J., Cabrita, M., Perramon, J., Munive, S., Guala, A., Kincl, V., Haitjema, S., Dantas, C., & Asselbergs, F. W. (2024). Requirements for human-centered artificial intelligence: A heart failure study across Europe and Latin America. *Proceedings of the International Symposium on Medical Information Processing and Analysis, SIPAIM*. <https://doi.org/10.1109/SIPAIM62974.2024.10783548>
- Roccas, S., Sagiv, L., Schwartz, S. H., & Knafo, A. (2002). The Big Five personality factors and personal values. *Personality and Social Psychology Bulletin*, *28*(6), 789–801. <https://doi.org/10.1177/0146167202289008>
- Rokeach, M. (1971). Long-range experimental modification of values, attitudes, and behavior. *American Psychologist*, *26*(5), 453–459. <https://doi.org/10.1037/h0031450>
- Rokeach, M. (1973). *The nature of human values*. Free Press.
- Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology* (Vol. 25, pp. 1–65). Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60281-6](https://doi.org/10.1016/S0065-2601(08)60281-6)
- Schwartz, S. H. (2012). An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, *2*(1). <https://doi.org/10.9707/2307-0919.1116>

- Shen, H., Knearem, T., Ghosh, R., Alkiek, K., Krishna, K., Liu, Y., Ma, Z., Petridis, S., Peng, Y.-H., Qiwei, L., Rakshit, S., Si, C., Xie, Y., Bigham, J. P., Bentley, F., Chai, J., Lipton, Z., Mei, Q., Mihalcea, R., ... Jurgens, D. (2024). *Towards bidirectional human-AI alignment: A systematic review for clarifications, framework, and future directions* (arXiv:2406.09264). arXiv. <https://doi.org/10.48550/arXiv.2406.09264>
- Shen, H., Knearem, T., Ghosh, R., Yang, Y.-J., Mitra, T., & Huang, Y. (2025). ValueCompass: A Framework of Fundamental Values for Human-AI Alignment. <https://doi.org/10.48550/arXiv.2409.09586>
- Simon, F., Nelson, R. K., & Fletcher, R. (2025). *Generative AI and news report 2025: How people think about AI's role in journalism and society*. Reuters Institute for the Study of Journalism. <https://reutersinstitute.politics.ox.ac.uk/generative-ai-and-news-report-2025-how-people-think-about-ais-role-journalism-and-society>
- Söderström, U., Hedström, E., Lambertsson, K., & Mejtøft, T. (2024). ChatGPT in education: Teachers' and students' views. *Proceedings of the European Conference on Cognitive Ergonomics 2024, ECCE '24*. <https://doi.org/10.1145/3673805.3673828>
- Sonni, A. F., Hafied, H., Irwanto, I., & Latuheru, R. (2024). Digital newsroom transformation: A systematic review of the impact of artificial intelligence on journalistic practices, news narratives, and ethical challenges. *Journalism and Media*, 5(4), 1554–1570. <https://doi.org/10.3390/journalmedia5040097>
- Tenório, K., & Romeike, R. (2024). AI Competencies for non-computer science students in undergraduate education: Towards a competency framework. *Proceedings of the 23rd Koli Calling International Conference on Computing Education Research, Koli Calling '23*. <https://doi.org/10.1145/3631802.3631830>
- UNESCO. (2019). *Beijing consensus on artificial intelligence and education*. United Nations Educational, Scientific and Cultural Organization. <https://unesdoc.unesco.org/ark:/48223/pf0000368303>
- UNESCO. (2025). *Youth survey report on AI and digital skills* (p. 28). https://unevoc.unesco.org/pub/wysd_survey_report_2025.pdf
- World Health Organization. (2024). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*. <https://www.who.int/publications/i/item/9789240084759>
- Yeste, V., & Rosso, P. (2026). *Human values in a single sentence: Moral presence, hierarchies, and transformer ensembles on the Schwartz continuum* (arXiv:2601.14172). arXiv. <https://doi.org/10.48550/arXiv.2601.14172>
- Yuan, P. H., Yan, T. D., Sharma, S., Chahley, E., MacLean, L. J., Freitas, V., & Yong-Hing, C. J. (2024). Authorship gender among articles about artificial intelligence in breast imaging. *European Journal of Radiology*, 175, Article 111428. <https://doi.org/10.1016/j.ejrad.2024.111428>