Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Semantic Alignment Between Normative Theories of Ethics and the European Union Artificial Intelligence Act: A Transformer-Based Semantic Textual Similarity Analysis

Mehmet Murat Albayrakoglu

20191006003@stu.khas.edu.tr

Kadir Has University

Mehmet Nafiz Aydin

Boğaziçi University

Article

Keywords: European Union Artificial Intelligence (EU AI) Act, Artificial Intelligence (AI) regulation, normative theory of ethics, Semantic Textual Similarity (STS), computational text analysis, AI governance

Posted Date: March 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8928758/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

The European Union Artificial Intelligence (EU AI) Act, which explicitly references fundamental rights and ethical principles, is a comprehensive regulatory framework for governing Artificial Intelligence (AI) systems. This study examines the moral grounding of the EU AI Act by analyzing the semantic alignment between three canonically distinct normative ethical theories (virtue ethics, deontological ethics, and consequentialism) and the Act's regulatory language. Building on philosophical and chronological considerations, the concept of influence is treated as a relational construct between the theories of ethics and the regulatory text. As a proxy for this relationship, Semantic Textual Similarity (STS) is employed to quantify the degree of alignment between the theory descriptions and the Act. The Act's preamble and statutory provisions are analyzed separately to capture its intentional and operational ethical groundings. To describe each theory distinctively and to reduce semantic overlap among theories, theory descriptions are manually preprocessed. To compute similarity scores, a heterogeneous embedding-level ensemble approach, comprising five lightweight Transformer-based encoders (SBERT, ALBERT, DistilBERT, RoBERTa, and TinyBERT), is used. To represent document-level alignment estimates, voting and averaging are used to aggregate STS scores. The findings indicate that deontological ethics exhibits the highest overall semantic alignment with both components of the EU AI Act.

Introduction

When legislation is devoid of ethical concern, deliberate or not, it may become an invitation to disaster because it lacks a moral basis to mitigate harm, again, intentional or not. Historically, many morally deficient laws have been enacted without due regard for ethical norms, and some have led to tragic consequences. (Radbruch (1946), 2006; Fuller, 1964). Among the examples are discriminatory legislation that favors a portion of society while condemning the rest, regulations that disregard potential environmental or health issues, laws that lead to the loss of historically and legally acquired property, and statutes that either disregard or outright violate fundamental human rights, or are unable to moderate individual and societal conflicts.

Information Technology (IT) legislation is no exception to ethical concerns, as there are examples of adverse outcomes resulting from the avoidance or disregard of the moral aspects of laws and regulations. Surveillance laws can lead to violations of fundamental rights (e.g., Goitein, 2019; Meireles, 2022) or the heavy-handed treatment of individuals and groups by governments (e.g., Human Rights Watch, 2019). Some statutes have weaknesses due to segmentation, which can exacerbate the difficulties faced by disadvantaged groups (e.g., McCarthy, 2004), and ambiguity, leading to indecisiveness or abuse (e.g., Kornbeck, 2021).

The European Union Artificial Intelligence (EU AI) Act has been subject to moral criticisms from the outset (e.g., Veale & Borgesius, 2021; Anderson, 2022). Despite being regarded as a significant step toward regulating Artificial Intelligence (AI) systems and emphasizing the fundamental rights (Musch et al., 2023), the risk-based approach taken to categorize and govern these systems is under scrutiny. One

criticism of the Act is its focus on domain-specific criteria and technical compliance, without considering power asymmetries, fairness, and autonomy (Veale & Borgesius, 2021).

This multidisciplinary work aims to examine the relationship between ethics and the EU AI Act. Its interdisciplinary nature was not solely derived from the inclusion of interpersonal and social influence, philosophy of ethics, and law. Among the contributing disciplines, computer science, and more specifically AI, was not only a subject of interest from ethical and legal perspectives, but it also provided the necessary tools and models to conduct analyses relevant to the study's aim.

The linguistic and computer science dimensions of the study were incorporated through the use of Natural Language Processing (NLP) and one of its subdivisions, Semantic Textual Similarity (STS), to demonstrate the existence and extent of the relationship between ethics and law. Due to resource constraints and environmental sensitivity, the study used lightweight Bidirectional Encoder Representations from Transformers (BERT) models to compute STS scores.

The central proposition of the study is: STS is a *proxy measure* of the influence of the composition and meaning of one textual document on another, given that, 1) there exists a common theme or context that encompasses both documents; and 2) preferably, a precedence, or at least a concurrency relationship over time, exists between the influencing document and the influenced document, respectively. At this point, STS is not claimed to establish causal influence, but to provide a comparative, text-related indicator of semantic alignment under explicit contextual and temporal conditions.

The influencing document is called the *influencer*, and the influenced document is called the *influencee*, a term we coined after the French *influenceé*, meaning influenced, and following the examples of the English words employee, lessee, and trainee, among others. In an influence relationship, the influencer precedes the influencee.

Influencers constitute the ethics dimension of the current work. For this study, three major, canonically distinct normative ethical theories were selected as influencers. They are virtue ethics, deontological ethics, and consequentialism. These theories are normative because they prescribe principles for morally acceptable attitudes, decisions, and actions and provide criteria for moral judgment.

The influencee is the EU AI Act of 2024, enacted to ensure that AI systems developed and used in the EU satisfy safety and transparency requirements without breaching ethical principles and fundamental rights. Within these confines, the Act allegedly aims to promote innovation and ensure the competitiveness of the EU institutions and businesses in AI.

Prior interdisciplinary research in information ethics, AI governance, and technology policy has examined how normative ethical principles shape the design, governance, and societal implications of computational systems and regulatory frameworks (e.g., Floridi, 2002; Capurro, 2006; Stahl, 2012; Floridi et al., 2018; Jobin et al., 2019; Veale & Borgesius, 2021). These works highlight that ethical concepts are often embedded, explicitly or implicitly, within socio-technical systems, governance mechanisms, and

legal discourse, particularly in the context of emerging AI regulation. Rather than engaging in prescriptive ethical judgment, such research frequently adopts analytical and interpretive approaches to examine how normative frameworks are reflected in institutional and policy texts.

Building on such an interdisciplinary perspective, the present study extends the discussion to regulatory language by computationally analyzing the semantic alignment between canonical normative theories of ethics and the European Union Artificial Intelligence Act, treating semantic similarity as a proxy indicator of potential influence rather than a direct measure of ethical or causal determination.

Textual documents in an influence relationship are manifestations of the thoughts and intentions of their creators. They may function as influencers or influencees depending on temporal precedence and information flow among them. Such relationships are socio-psychological and communicative in nature, as they involve the transmission of explicit or implicit meaning that can affect the beliefs, attitudes, or normative orientations of the influenced individuals. In this context, repetition or systematic alignment of compositional or semantic patterns between documents may signal influence, while the absence of such patterns suggests no discernible influence despite potential intent.

Because both the normative theories and the EU AI Act are textual artifacts reflecting the perspectives of philosophers and lawmakers, respectively, their semantic characteristics provide observable traces of this interaction. Given the historical depth and complexity of ethical traditions, constructing a detailed causal model of influence is impractical; however, established philosophical discourse on the ethics–law relationship supports the use of semantic textual similarity as a proxy measure for assessing the degree of influence between ethical theories and legal texts.

In the remainder of the paper, a literature review is presented that covers the key theoretical aspects of all contributing disciplines. Next, the contributions of each discipline are discussed and synthesized into a comprehensive methodology that employs a heterogeneous embedding-level ensemble approach. The approach uses five modified Bidirectional Encoder Representations from Transformers (BERT) models, built on the Transformer architecture, to calculate STS scores. These models are used to compare the theories pairwise with each of the two parts of the Act, namely, its preamble and statutory provisions, to calculate an STS score for each pair. The scores are sorted and averaged to determine which theory dominates. Finally, the model implementations are discussed, and conclusions are drawn about the work itself and its implications for the future.

Literature Review

The literature review examines prior work to contextualize the extent to which each normative theory relates to the EU AI Act. It covers the following: the relationship between ethics and law, the concept of interpersonal and social influence, the link between influence and semantics, and Semantic Textual Similarity (STS) as a proxy metric for the presence and extent of influence.

Moral Foundations of Law and Its Challenges to Natural Language Processing

Ethics of law, a branch of the philosophy of law, examines the relationship between ethics and law. Given the history of ethics and law, it is either impossible or prohibitively expensive to devise a detailed theoretical model that explicitly describes how ethics influenced law. In the absence of such a model, a qualitative approach appears more appropriate; the discourse among philosophers of law on the relationship between ethics and law provides the necessary insight into the problem.

Despite the division between the philosophers of law about the relationship between ethics and law, current work relies mainly on the views of legal antipositivists: the law should be motivated by the moral concerns of its stakeholders, that is, voters, legislators, public administrators, and judges (Slote, 2001; Jowitt, 2022). The law serves as a bridge between its moral foundations and legal institutions and their practices (Postema, 2022).

Even some legal positivists agree that ethics and law overlap, as both are based on norms aimed at preventing harm and promoting good, the former within individuals and the latter within a society (Kramer, 2004). Such a thematic relationship between ethics and law also forms the basis for their shared context: dos and don’ts to avoid harm and to do good according to conscience in the case of ethics and authority in the case of law.

In philosophical discourse, a single idea can be articulated through a diverse array of linguistic formulations (Rohatyn, 1972), and these introduce inherent ambiguities that complicate rigorous semantic analysis. This intrinsic variability in the language of philosophy (Adler & Doren, 1972; Gray, 2012; Martinich, 2016) implies that each expression, while conveying a core concept, may simultaneously introduce subtle shifts in meaning or emphasis. Consequently, using NLP techniques to analyze semantic relationships among philosophical texts, especially lengthy ones, creates considerable challenges (Jurafsky & Martin, 2025). To address such analytical difficulties and minimize their impact on the accuracy of meaning, the application of preprocessing techniques becomes a crucial methodological step.

Before delving into the methodology, however, a deeper analysis of the theoretical foundations of influence and semantics is necessary to develop a comprehensive conceptual framework for this study. Furthermore, it is crucial to unclutter the complexities arising from the inherent entanglement of influence relationships with semantic structures. Such a comprehensive understanding will provide clarity for subsequent analysis and interpretation.

The Concept of Influence and Its Relationship with Semantics

In a social setting, influence refers to the ability of one entity—an individual, group, or organization to affect the thoughts, beliefs, attitudes, decisions, or actions of one or more others in a specific manner

(Cialdini, 2021). In some cases, an influencer may use attraction, persuasion, or coercion to achieve desired outcomes (Tedeschi & Bonoma, 2017). In other cases, influence appears to be a spontaneous phenomenon. The difference between the two manifestations of influence is explained by compliance and conformity, respectively (Cialdini & Goldstein, 2004).

Influence is regarded as an outcome of some communication process (Back, 1951; Petty & Cacioppo, 1986). Back (1951) explored the dynamics of social influence within groups, laying a foundation for understanding broader group dynamics and the role of communication in influence processes. Petty and Cacioppo (1986) introduced the Elaboration Likelihood Model (ELM) to explain how people are persuaded and change their attitudes. ELM has been particularly influential in understanding how communication affects choice processes.

Influence as a communication process involves expression and interpretation. Therefore, there is an apparent connection between interpersonal or social influence and semantics. Semantics is a branch of linguistics that deals with the meaning of linguistic elements (Saeed, 2016; Qamar & Raza, 2024). It is not an exact science: among the problems of semantics are ambiguity and vagueness (Kennedy, 2019), which make it difficult to ascribe precise meaning to those linguistic elements.

In the past, a significant link between the concepts of influence and semantics has been established in the literature using three different approaches: theoretical, empirical, and AI-driven. One of the initial works (Halliday, 1978) used social semiotics to view language as a dynamic system in which meaning is constructed through the interaction of social structures and linguistic functions, involving ideational, interpersonal, and textual metafunctions.

In later theoretical treatments of influence-semantics relationships, Krauss and Fussell (1996) synthesized communication and cognitive models to emphasize the social and contextual nature of meaning construction. Saulwick and Trentelman (2014) formalized different types of influence using logical and linguistic constructs. Beltrama (2020) focused on social meaning, applying formal semantics and pragmatics to understand how linguistic forms convey information about users' social identities.

Empirical research examines how semantic cues shape beliefs, persuade people or groups, and alter social cognition. Gruenfeld and Wyer (1992) empirically studied how positive and negative statements shape beliefs and influence semantically related ideas. Fink et al.'s (2003) study provides a framework for differentiating persuasion from threats. Bargh et al. (2012) analyzed the role of automaticity in socio-cognitive processes, highlighting how the unconscious perception of others' behaviors and semantic associations can influence interpersonal behavior and social judgments.

Saint-Charles and Mongeau's (2018) study employed a socio-semantic approach. The authors analyzed meeting transcripts and sociometric data to examine the simultaneous evolution of social influence empirically and to identify discourse similarity within workgroups. Among more recent examples of empirical approaches, Jakesch et al. (2023) conducted a controlled experiment to analyze the impact of semantic suggestions on text generated by opinionated language models (such as GPT-3). Bian et al.

(2024) investigated the effect of external information on Large Language Models (LLMs) through a series of experiments. Breum et al. (2024) investigated the capacity of LLMs to shape opinions within synthetic social systems.

Within the AI-driven realm, Bayrakdar et al. (2020) examined the fundamental concepts of Social Network Analysis (SNA) by surveying various semantic analysis techniques applied to social media data (text, images, videos) to improve knowledge extraction and management. Goldstein et al. (2023) reported on the potential impact of LLMs on influence mechanisms. Finally, Bassi et al. (2024) surveyed and integrated classic persuasion theory with semantic modeling to study online persuasion.

The inherent link between influence and semantics suggests that the degree of semantic similarity or dissimilarity between textual expressions can be used to assess the extent of potential influence exerted or received. STS, a metric used to quantify the degree of semantic equivalence between a pair of texts (Bali et al., 2024), can be used to analyze and understand influence dynamics partially.

Semantic Textual Similarity

Despite the difficulty of attributing precise meaning to linguistic elements due to ambiguity and vagueness, measuring semantic similarity between two pieces of text remains a fundamental task in NLP (Zhao et al., 2024). By considering the lexical, syntactic, and semantic features of the texts, STS aims to quantify their similarity. Different authors have classified STS methods in different ways (Han et al., 2020; Wang & Dong, 2020; Chandrasekaran & Mago, 2021; He et al., 2024; Sasoko et al., 2024). Based on a combination of Han et al.'s (2020) and He et al.'s (2024) classifications, STS methods can generally be grouped into four major categories: String-based methods, corpus-based methods, knowledge-based methods, and deep-learning methods.

String-based methods focus on superficial features of two documents, such as word overlap, n-grams, or string matching. These methods take characters or words from both texts and return an STS score. They work well for duplicate detection and plagiarism analysis, but cannot handle synonyms, paraphrasing, or context shifts. They are computationally efficient but ignore the deeper semantic meaning of the texts. Despite their limitations, string-based approaches are valuable for baseline comparisons and are used during preprocessing stages in NLP pipelines.

Corpus-based methods utilize extensive collections of documents, known as corpora, to identify semantic relationships between word pairs. They aim to map words as vectors (word embeddings) in a high-dimensional space, positioning semantically similar words closer together. The STS score between a pair of texts is then derived from the similarity of their respective word embeddings. These methods rely on pre-trained embeddings, such as Word2Vec and GloVe, and often employ similarity metrics, including cosine similarity and Euclidean distance, to quantify semantic closeness. Thus, analyzing linguistic patterns across large datasets helps capture nuances in semantic relationships beyond lexical similarity.

Knowledge-based methods use predefined relationships between words and concepts to assess semantic similarity. They use structured linguistic resources, such as ontologies, lexical databases, and semantic networks, to establish semantic relationships between words. Then, they aggregate these similarities to yield an STS score. These methods are particularly valuable because they provide interpretability and domain specificity, which are often lacking in corpus-based models due to data sparsity. However, they struggle with scalability and lack the flexibility to adapt to new linguistic variations beyond predefined taxonomies.

Deep learning methods use neural networks to learn semantic representations of text. Unlike corpus-based or knowledge-based approaches, these methods rely on hierarchical feature extraction and contextual embeddings from deep architectures such as Transformers and recurrent networks. They can effectively capture contextual and semantic relationships and directly model the similarity between sentences or passages. Deep-learning STS approaches surpass traditional similarity metrics because they can capture nuanced language relationships, including paraphrasing and implicit meanings. However, they require substantial computational resources and large-volume corpora annotated for effective generalization.

The specific method for obtaining an STS score depends on two main factors: the texts to be compared and the available resources. These two factors can be further refined across the categories of STS methods.

String-based models and corpus-based approaches are transparent and fast but may not capture semantic nuances. A trade-off between explainability and semantic depth is necessary in selecting an STS method (Manning & Schütze, 1999). Corpus and knowledge-based STS methods are suitable in low-resource environments. However, it is essential to balance semantic richness, data scarcity, and computational constraints (Mihalcea et al., 2006).

Agirre et al. (2012) introduced two additional factors: generalizability and human judgment. Generalizability refers to an STS method's ability to maintain reliable performance across different types of texts, domains, and languages. According to the authors, an STS method should consistently yield STS scores that align with human intuition, regardless of the domain in which it is applied. Cer et al. (2017) extended these criteria to deep-learning methods over six genres: news, forums, headlines, image captions, and question-answer pairs, in addition to corpus-based methods.

In recent years, deep-learning models have taken the lead in STS research, shifting the focus to selecting and applying the most effective method. Devlin et al. (2019) presented Bidirectional Encoder Representations from Transformers (BERT) to enhance language understanding. Vaswani et al. (2017) proposed the Transformer architecture based on self-attention, which served as a precursor to BERT. Self-attention is a mechanism where each element in a word sequence computes a weighted sum of all elements in that same sequence. Learned similarity scores between elements determine their weights. Employing self-attention mechanisms enables parallel computations. Thus, self-attention significantly improves efficiency in NLP tasks that require deep context understanding.

BERT leverages deep bidirectional attention by simultaneously considering both left- and right-textual contexts of a string. It enhances NLP task performance by employing a pre-training strategy that combines Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). MLM is based on random masking and the prediction of words. NSP helps the model identify relationships between sentences. These pre-training strategies enable more precise semantic representations and improve contextual understanding. With minimal modifications, developers fine-tuned BERT for tasks such as text classification, named entity recognition, and question answering.

Reimers & Gurevych (2019) modified BERT, called Sentence-BERT (SBERT), which is based on pairwise comparisons. By employing such comparisons, SBERT overcomes the computationally expensive large-scale similarity searches typically required by traditional BERT models. STS scores are calculated using cosine similarity within a fixed-size vector space.

Similarly, Lan et al. (2019) proposed A Lite BERT (ALBERT) to address the limitations of the original BERT model without compromising performance. The authors employed two techniques, factorized embedding parameterization and cross-layer parameter sharing, to reduce resource requirements and to improve performance in multi-sentence understanding for longer texts.

Another low-compute Transformer model, DistilBERT, was presented by Sanh et al. (2019). The model is based on knowledge distillation to develop a smaller general-purpose language representation. Their approach can be fine-tuned for various tasks with only a slight performance sacrifice.

Liu et al. (2019) introduced an improved version of BERT, known as the Robustly Optimized BERT Pretraining Approach (RoBERTa). The authors removed the Next Sentence Prediction (NSP) from the original model. They trained it on longer sequences than those used in BERT and employed dynamic masking to enhance the accuracy and performance of their model. Dynamic masking enables the model to learn from multiple masking patterns per sentence, thereby adapting to various sentence structures. Thus, they required fewer resources than BERT, which uses whole-word masking.

Finally, Jiao et al. (2019) introduced TinyBERT, a compact and efficient variant of BERT designed to reduce computational requirements while preserving performance. TinyBERT employs a two-stage knowledge distillation framework, comprising pretraining and fine-tuning. As a result, the model is substantially smaller and faster than BERT while achieving comparable performance on NLP tasks.

While presenting their models, the developers of the five lightweight BERT models also mentioned their criteria for selecting the most suitable model, either explicitly or implicitly. Table 1 summarizes the criteria they stated or implied.

This study employs these five lightweight transformer-based models—SBERT, ALBERT, DistilBERT, RoBERTa, and TinyBERT—as sentence encoders to compute semantic textual similarity (STS). The following section details the research design, preprocessing procedures, and model application steps used to operationalize this approach.

Table 1
Model selection criteria for the lightweight BERT models

Model	Dominant Selection Criteria
BERT (Devlin et al., 2019)	Context-sensitive representations; accuracy prioritized over efficiency
SBERT (Reimers & Gurevych, 2019)	Sentence-level semantics; low-resource and real-time suitability
ALBERT (Lan et al., 2019)	Memory and parameter efficiency; scalable similarity scoring
DistilBERT (Sanh et al., 2019)	Reduced model size with acceptable accuracy
RoBERTa (Liu et al., 2019)	Performance optimization within existing criteria; accuracy maximization through improved training
TinyBERT (Jiao et al., 2019)	Extreme compression; accuracy-cost trade-offs

Methodology

Figure 1 summarizes the study's research design and analytical workflow. Three canonically distinct normative theories of ethics are selected as influencer texts. At the same time, the EU AI Act is partitioned into its preamble and statutory provisions to distinguish its intentional and operational aspects. Following high-level text preprocessing to minimize semantic overlap, semantic textual similarity (STS) scores are computed using a heterogeneous ensemble of lightweight transformer-based models. Pairwise comparisons are performed between each theory of ethics and each component of the Act, yielding sentence-level similarity scores that are further aggregated to the document level. The resulting scores are analyzed and visualized to assess relative patterns of ethical alignment across models.

The Data: Three Normative Theories of Ethics and the EU AI Act

This subsection summarizes the influencing documents—encyclopedic treatments of the influencers, virtue ethics, deontological ethics, and consequentialism—and the influencee, the EU AI Act. Although scholars wrote them, there are several reasons for choosing encyclopedic entries as influencers over original philosophical treatments, scholarly books, articles, or textbook narratives. First, encyclopedia entries help avoid subjective or argumentative narratives, which are likely to add complexity to the

machine's understanding of texts. Second, the materials should not be targeted at a specific segment of the audience. Third, the degree of textual structure of the influencers and the influencee should be as compatible as possible. Fourth, a consistent, up-to-date terminology and vocabulary should be used in narratives drawn from a range of philosophical resources, including translations and historical works.

The first normative theory, *virtue ethics* (Hursthouse & Pettigrove, 2023), emphasizes the importance of virtues and a person's moral character in action. The ethical decisions and actions of a virtuous person, who strives to do what is right, good, just, or proper, are based on the person's character and their actions. Virtue ethics is based on three ideas from ancient Greek philosophy: *phronesis* (moral or practical wisdom), *eudaemonia* (happiness or flourishing), and *arete* (excellence or virtue). Virtues are admirable character traits that guide a person's attitudes and actions. An ethical person is honest, wise, fair, courageous, and self-controlled. However, virtue is a matter of degree; perfect or flawless virtue is uncommon.

The second normative theory, *deontological ethics* (Alexander & Moore, 2021), categorizes actions as morally required, forbidden, or permitted. It guides and assesses a person’s choices of what they ought to do. Deontological approaches hold that some options are morally forbidden even if their overall effect would be good. Deontologists believe that a choice is right if it conforms to a moral norm that each moral agent should obey. Deontological ethics is founded on the following three rules: 1) Do what you would want to be done to you, by others, and to others; 2) Always apply the same rules to everybody, including yourself; and 3) a person is never a means but an end for themselves. Some deontologists focus on agency and the idea that morality is, to some extent, a personal matter.

The final normative theory, *consequentialism* (Sinnott-Armstrong, 2023), suggests that moral rightness or wrongness depends solely on the consequences of one’s decisions and actions. *Hedonism* holds that pleasure is the only intrinsic good and pain the only intrinsic evil. Classic utilitarians hold a hedonistic act-consequentialist view, which claims that an act is morally justified if it causes the greatest happiness for the greatest number of stakeholders. Additional normative characteristics included in consequentialist theories should depend solely on consequences. There are several shades of consequentialist theories, such as maximizing consequentialism, hedonistic consequentialism, and aggregative consequentialism. What distinguishes one view from another is the extent to which moral rules are included.

These theories have some commonalities. First, any description of a theory of ethics should be centered around ethically acceptable or unacceptable attitudes, decisions, or acts. Ethical behavior lies on a continuum between totally acceptable or positive, and unacceptable or negative. On the positive end, human acts are often qualified as right, good, fair, virtuous, appropriate, beneficial, impartial, and unbiased, among other terms. On the negative end, they are wrong, bad, unfair, vicious, inappropriate, harmful, partial, biased, etc.

Apart from this central theme, there are other similar aspects in the theories of ethics. Table 2 qualitatively lists the most significant similarities between each pair of the three major normative

theories of ethics, derived from pairwise comparisons of the preprocessed entries from the Stanford Encyclopedia of Philosophy (Alexander & Moore, 2021; Hursthouse & Pettigrove, 2023; Sinnott-Armstrong, 2023), and the works of Kagan (1998) and Wood (2020).

On the table, all three theories claim to be normative and universal, while also taking circumstances into account and acknowledging multiple perspectives on a good life. Additionally, each theory rejects pure forms of the others while maintaining normativity and universality, suggesting that these frameworks are complementary rather than mutually exclusive.

There are also pairwise similarities between the theories. Virtue and deontological ethics emphasize practical reason, moral excellence, character, duty-based thinking, and the priority of right over good, each to a significant, yet varying degree. Deontological ethics and consequentialism share an emphasis on moral reasoning, impartiality, stakeholder relations, consideration of circumstances, the importance of intention, and the priority of good over mere praiseworthiness. Finally, deontological ethics and consequentialism together highlight moral reasoning, impartiality, and stakeholder relations, consideration of circumstances, importance of intention, and the priority of good over mere praiseworthiness.

Table 2
A qualitative summary of similarities between pairs of the three normative theories of ethics

Virtue Ethics and Deontological Ethics	**Virtue Ethics and Consequentialism**	**Deontological Ethics and Consequentialism**
Normativeness	Normativeness	Normativeness
Universality	Universality	Universality
Priority of right over good	Multiple views of a good life	Multiple views of a good life
Emphasis on moral excellence	Presence of idealized moral agents	Emphasis on moral reasoning
Importance of practical reason	Consideration of circumstances	Consideration of circumstances
Importance of obligations	Importance of happiness	Emphasis on impartiality
Importance of character	Importance of human relationships	Emphasis on stakeholder relations
Importance of intention	Rejection of pure deontology	Priority of good over praiseworthy
Rejection of pure utility	Rejection of pure emotions	Importance of intention
	Rejection of pure rationality	

The EU AI Act (European Parliament & Council of the European Union, 2024) regulates the development and use of AI systems across EU member states. It also applies to non-EU companies operating inside

the EU. The Act's objectives are to ensure the safe use of AI systems, support fundamental rights, and foster AI innovation within the EU.

The Act defines four categories of risk for AI systems. The highest level is *unacceptable risk*, which refers to AI systems and practices regarded as harmful or unethical. Such systems threaten fundamental EU values and rights, and, consequently, the Union prohibits their use across its member states. *High-risk* systems are those used in mission-critical sectors, such as healthcare and law enforcement, or those that may compromise fundamental rights. These systems are subject to rigorous controls and oversight. *Limited-risk* systems can be used for deception and manipulation and require specific transparency measures. *Minimal-risk* systems pose almost no risk to individuals' safety or fundamental rights. They are not subject to any particular regulatory obligation.

In addition to risk categories, the EU AI Act emphasizes the fundamental rights of EU citizens, including human dignity, freedom, equality, and democracy. It also encourages the rule of law and respect for human rights. The Act distinguishes between the developers and users of AI systems and specifies several obligations for both groups for high-risk systems.

Due to the increasing capabilities and potential impact of recent AI systems, such as Generative AI (GenAI) applications, the Act introduces specific transparency requirements for their developers, regardless of their intended purpose of use. As these applications become more powerful, they are subject to additional, stricter requirements for model evaluation, risk assessment and mitigation, incident reporting, and cybersecurity.

The law also introduces a robust governance and enforcement framework, including the establishment of an EU AI Office within the European Commission (EC) and the requirement for member states to create national AI offices. It sets significant penalties for non-compliance, depending on the severity of the infringement and the size of the EU and non-EU developers or users of the systems. Finally, the Act establishes provisions for scope changes, recognizing the dynamic nature of AI technology.

Justification of the Fundamental Proposition

In the literature review, it has already been established that semantic similarity can be used to measure the influence of one textual document on another, provided that a shared context and a precedence relationship exist between the influencer and the influencee. In the literature review, the discussion of the ethics of law has already demonstrated the mutual context between ethics and law.

To establish the precedence relationship, the years of publication of each major official EU AI Act document should be compared to the years of publication of the resources collected and consulted in the bibliography for each theory of ethics described.

The first initiative leading to the EU AI Act started with the Consultation on Artificial Intelligence, launched in February 2020. The results were published in a white paper titled “Public Consultation on the

AI White Paper: Final Report" in November 2020 (Directorate-General for Communications Networks, Content and Technology, 2020). In April 2021, the proposal was presented to the European Parliament by the Commission's Directorate-General for Communications Networks, Content, and Technology (2021). Finally, it was enacted by the European Parliament in March 2024, approved by the European Union Council in May 2024, and came into force on August 1, 2024, with some provisions covering up to 3 years after the enforcement of the law (European Parliament & Council of the European Union, 2024).

The descriptions of the theories of ethics are taken from the Stanford Encyclopedia of Philosophy (Zalta & Nodelman, n.d.). The bibliography of the entry Virtue Ethics spans the period between 1956 and 2021 (Hursthouse & Pettigrove, 2023), although its roots can be traced back to Plato (429 BC-347 BC) and Aristotle (384 BC-322 BC) (Van Zyl, 2019). The bibliography of Deontological Ethics spans the 18th century to 2019 (Alexander & Moore, 2021). Likewise, Consequentialism's bibliography starts in 1755 and ends in 2020 (Sinnott-Armstrong, 2023).

Compared to the period covered by the EU AI Act documentation, the bibliographies of the normative theories of ethics originated from works written centuries earlier. Consequently, all three theories precede the Act, and the requirement for the precedence relationship is satisfied.

Semantic Textual Similarity (STS) with Lightweight BERT Models

This study uses five models defined over the semantic space and discussed in the literature review to calculate STS. Semantic space is considered a normalized metric space in which distance is used to measure semantic similarity (Rozinek & Mareš, 2021). Distinct from the lexical semantics that apply to words, STS applies to larger units of language: sentences, paragraphs, or longer pieces consisting of multiple paragraphs, sections, chapters, parts, or entire textual artifacts. The term "metric" refers to the measurement of the distance between two points in space (Dshalalow, 2013).

Lightweight BERT variants utilize *cosine* similarity, which emphasizes direction over distance, thereby enhancing the model's performance. Since it does not directly satisfy the metric space axioms, cosine similarity is referred to as a *pseudo-metric*. Nevertheless, it can be converted to a cosine distance through an $arccosine$ transformation, and cosine distance is a metric.

A term-frequency vector, consisting of the number of occurrences of each term in a document, represents the document. Similarity between two documents is calculated by applying the following formula to their vectors:

$$\sigma\left(\mathrm{varvec}x, \mathrm{varvec}y\right) = \frac{\mathrm{varvec}x \bullet \mathrm{varvec}y}{\|\mathrm{varvec}x\|\|\mathrm{varvec}y\|} \quad \text{(Eq. 1)}$$

where

$\sigma\left(\mathrm{varvec}x, \mathrm{varvec}y\right)$ = Similarity of two term-frequency vectors $\mathrm{varvec}x$ and $\mathrm{varvec}y$

$\varvec{x}$ = term-frequency vector of document x

$\varvec{y}$ = term-frequency vector of document y

$||\varvec{x}||$ = Euclidean norm of the vector $\varvec{x}$

$||\varvec{y}||$ = Euclidean norm of the vector $\varvec{y}$

$\sigma\left(\varvec{x}, \varvec{y}\right)$ is a measure of how close two non-zero vectors are in an inner product space. The closer the pair of vectors is, the more similar the two documents (Han et al., 2012).

Text Preprocessing

Text preprocessing is crucial in NLP because raw text often contains noise, inconsistencies, and redundancies that can negatively impact model performance. In Semantic Textual Similarity (STS), preprocessing ensures that models accurately capture meaning rather than surface-level differences. In the STS literature, however, text processing refers to operations such as tokenization, lowercasing, stopword removal, and stemming applied to the lexical elements of a text (Chai, 2023).

To avoid confusion about what each theory is about, the higher-level preprocessing rules shown in Table 3 were uniformly applied to the sentences, paragraphs, and words in the descriptions of the theories of ethics. The purpose of using these rules is to minimize semantic overlap among the descriptions of theories. These rules help eliminate linguistic elements that could interfere with the meaning of descriptions, as they are more relevant to another theory from a machine learning perspective. Additionally, each rule was justified by adding a rationale immediately following it. In applying these rules, care was taken to ensure that no rule altered evaluative content, normative claims, or core vocabulary.

The text of the Act itself is divided into two parts, the preamble and provisions, to ascertain if these two parts are theoretically consistent from an ethical point of view. However, both parts are used as-is without modification.

The Ensemble Approach

In NLP, embedding-level and multi-encoder ensembles are applied to semantic similarity. Among the models used in this research, SBERT is a sentence-embedding model by design. ALBERT, DistilBERT, RoBERTa, and TinyBERT are token-level transformer encoders that can be adapted to produce sentence embeddings via pooling strategies, such as mean pooling or classification-based (CLS) pooling.

Table 3
Text preprocessing rules applied to theories of ethics

Rule No.	Description	Rationale
1	Remove titles, subtitles, etc.	They do not describe a theory but indicate specific parts of the document.
2	Eliminate meta descriptions (descriptions of the document, e.g., TOC, abstract, etc.)	They do not describe the theory; instead, they show what the document is about or how it is organized.
3	Remove items from the reference list.	They do not describe a theory.
4	Remove proper nouns.	Not the nouns but the ones they belonged to described each theory.
5	Delete discussions about what the theory is not, but keep negative examples.	Those discussions do not describe a theory.
6	Keep only the conclusive statements for the incremental arguments.	Eliminate irrelevant words.
7	Keep comparisons of the various forms of the same theory.	They are indispensable extensions of a theory.
8	Remove references to religions and religious symbols.	Ensure religious neutrality.
9	Delete descriptions of, references to, or comparisons with other theories.	Isolate one theory from another to preserve context.
10	Convert text into US English.	Eliminate variations in spelling.
11	Replace foreign-language words (mostly Greek or Latin) with their US-English equivalents, if any (Collins and Merriam-Webster).	Reduce the likelihood of encountering missing words in the model.
12	Add English translations of foreign-language words if there is no US-English equivalent.	Reduce the likelihood of encountering missing words in the model.

The application of ensemble methods in text mining involves adapting different algorithms and models to validate the results (Zong et al., 2021). We employed an embedding-level ensemble approach to balance semantic expressiveness with computational efficiency and methodological reproducibility. Independent STS scores yielded by each lightweight BERT encoder were aggregated to improve robustness and reduce model-specific biases (Dietterich, 2000; Zhou, 2012). The ensemble approach enabled us to capture complementary semantic representations and to produce more stable and reliable similarity estimates than those from a single model.

Results

Once the text preprocessing was complete, each theory, along with the Act's preamble and then with the provisions, was fed into each of the five lightweight BERT models identified in the literature review, and STS scores were recorded. Tables 4 (a) and (b) summarize the STS scores for the Act's preamble and provisions, respectively. The column captioned "Model Identifier" on the table refers to the specific pre-trained version of the Transformer model to the left of it. Except for the TinyBERT model, Table 4 shows that deontological ethics influence both the preamble and the provisions more than virtue ethics and consequentialism do. Therefore, on average, deontological ethics dominates the other two. Although the rank of consequentialism's influence varies across the models, on average, it comes second, and virtue ethics, again with variations in ranking, has the least impact.

Table 4
STS scores between each theory of ethics and the two parts of the EU AI Act(a) STS scores between each theory of ethics and the preamble of the EU AI Act.

		STS of Theories and the Act's Preamble (%)		
Transformer Model	**Model Identifier**	**Virtue Ethics**	**Deontological Ethics**	**Consequentialism**
SBERT	all-MPNet-base-v2	18.80%	26.62%	11.73%
ALBERT	paraphrase-albert-small-v2	15.05%	21.14%	18.09%
DistilBERT	distilbert-base-nli-stsb-mean-tokens	36.13%	40.30%	36.76%
RoBERTa	all-distilroberta-v1	14.34%	21.36%	15.88%
TinyBERT	paraphrase-TinyBERT-L6-v2	17.32%	14.81%	26.48%
	Average	20.33%	24.85%	21.79%
	Maximum	36.13%	40.30%	36.76%
	Minimum	14.34%	14.81%	11.73%
	Range	21.79%	25.49%	25.03%

(b) STS scores between each theory of ethics and the provisions of the EU AI Act

		STS of Theories and the Act's Provisions (%)		
Transformer Model	**Model Identifier**	**Virtue Ethics**	**Deontological Ethics**	**Consequentialism**
SBERT	all-MPNet-base-v2	9.92%	20.61%	2.26%
ALBERT	paraphrase-albert-small-v2	12.40%	19.81%	15.61%
DistilBERT	distilbert-base-nli-stsb-mean-tokens	36.57%	42.29%	39.41%
RoBERTa	all-distilroberta-v1	14.08%	18.54%	16.72%
TinyBERT	paraphrase-TinyBERT-L6-v2	19.13%	14.67%	26.53%
	Average	18.42%	23.18%	20.11%
	Maximum	36.57%	42.29%	39.41%
	Minimum	9.92%	14.67%	2.26%

		STS of Theories and the Act's Preamble (%)		
Transformer Model	**Model Identifier**	**Virtue Ethics**	**Deontological Ethics**	**Consequentialism**
	Range	26.65%	27.62%	37.15%

Figures 2 (a) and (b) provide further insight into the results of Table 4 (a) and (b), respectively. The most notable is the considerably higher STS scores obtained with the DistilBERT model compared to the other models. Scores other than DistilBERT's are accumulated toward the center of the radar chart. Another significant finding is the high variation in the STS scores supplied by the SBERT model across the three theories. This is especially noticeable for the STS scores of the Act's provisions in Fig. 2 (b). The consistently higher similarity scores produced by DistilBERT likely reflect architectural or training-specific embedding properties rather than substantive ethical alignment. However, we cannot be certain, but we can only speculate about the causes of DistilBERT scores due to the opacity of Transformer models.

Until now, we have overlooked the possibility of interrelationships, whether influential or not, between pairs of theories of ethics. Therefore, the assumption underlying the results presented in Table 4 and Fig. 2 is that either no semantic relationships exist among the three theories or, if they do, the interactions are negligible. This point requires further elaboration in light of Table 2, as STS should be considered a measure of influence not only between the theories and law, but also for the lateral semantic interactions among the theories themselves.

In Table 5, the pairwise similarities of the theories shown in Table 2 are quantified by utilizing the same lightweight BERT models to show how closely their textual descriptions align with each other. On the table, deontological ethics and consequentialism show the strongest textual similarity. Virtue ethics and deontological ethics, as well as virtue ethics and consequentialism, share moderate similarities.

Figure 3 graphically illustrates the findings presented in Table 5, providing further insight into the pairwise STS scores. The STS scores for the ALBERT, DistilBERT, and RoBERTa models indicate that the preprocessed descriptions of deontological ethics and consequentialism are semantically most similar. In contrast, the SBERT model suggests that the descriptions of virtue ethics and deontological ethics are the most similar. The TinyBERT model suggests that virtue ethics and consequentialism are the most similar.

However, when explaining these findings, the models raise new questions rather than provide explanations, since the specifics of calculating each STS score are unknown. In contrast, the principles (algorithms), datasets, and identifiers employed by each model are known.

Table 5
STS scores between pairs of the three nominal theories of ethics

Transformer Model	STS (%)		
	Virtue Ethics and Deontological Ethics	**Virtue Ethics and Consequentialism**	**Deontological Ethics and Consequentialism**
SBERT	44.12%	34.67%	41.96%
ALBERT	30.80%	33.48%	36.21%
DistilBERT	56.31%	56.60%	57.87%
RoBERTa	39.83%	34.03%	47.93%
TinyBERT	30.24%	33.27%	30.48%
Average	40.26%	38.41%	42.89%
Maximum	56.31%	56.60%	57.87%
Minimum	30.24%	33.27%	30.48%
Range	26.07%	23.33%	27.39%

Discussion

The results from Table 4 and Fig. 2 reveal a consistent pattern: Deontological ethics shows the highest semantic similarity to both the EU AI Act's preamble and statutory provisions, followed by consequentialism and virtue ethics. The low average similarity scores (around 25%) suggest that none of the three theories of ethics has had a sole influence on the EU AI Act. Instead, these results suggest that AI governance may be evolving into a novel form of "regulatory ethics" that selectively incorporates elements from multiple philosophical traditions to address the unique challenges posed by AI technologies.

The consistent dominance of deontological ethics across both the preamble and provisions deserves deeper analysis. Deontological ethics is fundamentally rule-based, emphasizing duties, rights, and categorical principles. So are the legal texts. However, this raises an interpretive question: Does the higher similarity reflect genuine philosophical alignment with Kantian principles of human dignity and categorical imperatives or merely structural similarities between rule-based ethical systems and legal language? The EU AI Act's focus on stakeholder rights, banned practices, and compliance rules can create language patterns that align with duty-based deontological ethics, regardless of the underlying moral principles.

Future research should compare the EU AI Act with other legal texts to distinguish between structural and substantive similarities and establish a baseline for the similarity in legal language. Furthermore, the language of specific rights and duties should be freed from general rule-making structures, and the Act's stakeholder protections should be examined to determine whether they reflect genuine deontological principles or merely procedural compliance requirements.

The limited corpus of texts representing each theory of ethics is a significant constraint. Each philosophical tradition spans centuries and encompasses multiple schools of thought that might yield different similarity scores.

The preprocessing of legal texts presents unique challenges that may have affected similarity calculations. Legal documents employ formatting conventions (article numbers, paragraph structures, cross-references) and syntactic patterns (dependent clauses, conditional statements, definitional sections). Treating complex legal sentences with multiple dependent clauses as separate units breaks down the logical structure of the legal reasoning. While necessary for computational processing, this approach may obscure the integrated argumentative patterns that characterize legal and philosophical discourse.

Table 5's revelation that inter-theory STS scores significantly exceed theory-to-AI Act scores creates another interpretive puzzle. This pattern suggests that the AI Act covers a wider ethical territory, encompassing traditional philosophical categories. It may also indicate a significant difference between legal language and philosophical discourse, leading to dissimilarities among related concepts. It may even be a symptom of poorly calibrated semantic similarity metrics for analyzing philosophical content.

Current methodologies cannot incorporate the inter-theoretical semantic similarities presented in Table 5 into the STS scores in Table 4, as they assume independent comparison units. However, philosophical theories exist in discursive relationships with one another, influencing legal frameworks through recursive processes in which ideas are repeatedly combined, critiqued, and modified for practical use.

Despite these methodological challenges, the findings offer valuable insights into the ethical foundations of AI governance. The modestly consistent alignment between deontological ethics and the EU AI Act suggests that rights-based, duty-oriented approaches to AI ethics may have gained prominence in regulatory contexts. This finding has practical implications that depend on the stakeholders' power and attitudes. On the one hand, if AI regulation leans toward deontological frameworks, we might expect future governance approaches to emphasize categorical restrictions, inviolable rights, and duty-based compliance obligations. On the other hand, if it deviates from deontological frameworks, we might expect future governance approaches to emphasize consequentialist cost-benefit analyses, or, to a lesser extent, virtue-based professional ethics standards.

This discussion reveals that while STS analysis opens new possibilities for investigating philosophical influences on policy, it also generates interpretive challenges that require careful methodological

development and theoretical sophistication to produce more specific results.

Conclusion

This study aims to utilize NLP to gain insight into the interaction between the philosophy of ethics and law, using a limited number of texts and a limited number of Transformer-based models. The results suggest that deontological ethics exhibits the highest semantic alignment with the EU AI Act among the theories examined. This may be interpreted as a stronger proxy indicator of potential ethical influence.

However, since almost all the STS scores per model are of the same order of magnitude, we can safely conclude that the others exerted lesser but comparable influences.

This study should be regarded as a first attempt, and its caveats should be noted, as the number of texts on each theory needs to be increased to assess the degree of agreement on which theory dominates the Act. Therefore, several iterations of the work with different yet comparable accounts of the theories would yield more sound results.

Considering the Act, one potential issue is using its text as is. Whether keeping titles, paragraph numbers, or article numbers has distorted the results is unknown. The Act includes very long sentences, especially complex ones, in which each of the multiple dependent clauses appears as a distinct statement, each starting on a new line. This presents a challenge that may need to be addressed: repeating the independent clauses before each of their dependents does not appear to be a viable solution.

Aside from being compared to human judgment, the lightweight BERT models in this study lack transparency, and none can justify their STS scores. Although they can yield STS scores for intertheoretical influences, these scores cannot be justifiably incorporated into the STS scores of theories and the Act. Apparently, there is a need for a new approach to account for the lateral interactions among influencers and properly combine them into the STS scores for theories and law.

Future research in this area requires development along several dimensions: First, the corpus should be extended by incorporating larger, more representative collections of texts of ethics that capture the diversity within each theoretical tradition while maintaining comparability across theories. Second, transparent similarity metrics should be developed to identify the specific textual features that drive philosophical alignment, including vocabulary, argumentative structure, normative claims, and prescriptive language patterns. Third, dynamic modeling approaches represent perhaps the most important methodological frontier. There is an obvious need for feasible dynamic models that can account for the discursive relationships between philosophical theories and their combined influence on legal texts, rather than treating each theory as an independent influence factor.

For the time being, this work’s theoretical and methodological contribution to the literature surpasses its findings. Our work demonstrated NLP's ability to help uncover influence relationships that may have

been lost over time or in translation. Even though our approach cannot be used to prove the presence of influence relationships, it can be used either to support and strengthen the arguments for their existence, as in the case of ethics and law handled in this work; or to develop hypotheses about influence relationships that may lead to further research, not only in ethics, but also in history, sociology, law, politics, literature, arts, etc.

Declarations

Ethical Approval and Informed Consent:

This article does not contain any studies with human participants performed by any of the authors.

Funding

Declaration

No funding was received for this study.

Author Contribution

M.M.A. conceived the study, developed the methodology, conducted the analysis, and drafted the manuscript. M.N.A. provided conceptual guidance throughout the research process, offered critical feedback on the manuscript, and contributed to refining the study's arguments. Both authors reviewed and approved the final manuscript.

Data Availability

All data, and Python scripts, as well as information about data processing, computer interpretation environment, and the models used in our study were included in a ZIP file, named "hssc_data_scripts_ethics_reqs.zip." The file was submitted as supplementary material. Since the data we used is copyrighted, it is for the information of the editors and referees only.

References

1. Adler, M. J. & Doren, C. (1972). *How to read a book* (Rev. ed.). Simon and Schuster.
2. Agirre, E., Cer, D., Diab, M., & Gonzalez-Agirre, A. (2012). SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (SEM)*, (pp. 385–393). Association for Computational Linguistics.
3. Alexander, L., & Moore, M. (2021). Deontological ethics, *The Stanford Encyclopedia of Philosophy, Winter 2021 Ed*. Zalta, E. N. (ed.). Retrieved June 14, 2024, from

https://plato.stanford.edu/archives/win2021/entries/ethics-deontological.

4. Anderson, M. M. (2022). Some Ethical Reflections on the EU AI Act. *1st International Workshop on Imagining the AI Landscape After the AI Act (IAIL 2022)*. CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3221/IAIL_paper5.pdf
5. Back, K. W. (1951). Influence through social communication. *The Journal of Abnormal and Social Psychology, 46*(1), 9–23.
6. Bali, A., Bhagwat, A., Bhise, A., & Joshi, S. (2024). Semantic similarity detection and analysis for text documents. In *2024, the 2nd International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)*, (pp. 1–9). IEEE. http://dx.doi.org/10.1109/ic-ETITE58242.2024.10493834
7. Bargh, J. A., Schwader, K. L., Hailey, S. E., Dyer, R. L., & Boothby, E. J. (2012). Automaticity in social-cognitive processes. *Trends in Cognitive Sciences, 16*(12), 593–605. https://doi.org/10.1016/j.tics.2012.10.002
8. Bassi, D., Fomsgaard, S., & Pereira-Fariña, M. (2024). Decoding persuasion: A survey on ML and NLP methods for the study of online persuasion. *Frontiers in Communication, 9*, 1457433.
9. Bayrakdar, S., Dogru, I. A., Yucedag, I., & Simsek, M. (2020). Semantic analysis on social networks: A survey. *International Journal of Communication Systems, 33*(e4424). http://dx.doi.org/10.1002/dac.4424
10. Beltrama, A. (2020). Social meaning in semantics and pragmatics. *Language and Linguistics Compass, 14*(10), e12398. https://doi.org/10.1111/lnc3.12398
11. Bian, N., Lin, H., Liu, P., Lu, Y., Zhang, C., He, B., Han, X., & Sun, L. (2024). Influence of external information on large language models mirrors social cognitive patterns. *IEEE Transactions on Computational Social Systems* (Early Access), 1–17. https://doi.org/10.1109/TCSS.2024.3476030
12. Breum, S. M., Egdal, D. V., Gram Mortensen, V., Møller, A. G., & Aiello, L. M. (2024). The Persuasive Power of Large Language Models. *Proceedings of the International AAAI Conference on Web and Social Media*, *18*(1), 152–163. https://doi.org/10.1609/icwsm.v18i1.31304
13. Capurro, R. (2006). Towards an ontological foundation of information ethics. *Ethics and Information Technology, 8*(4), 175–186. https://doi.org/10.1007/s10676-006-9108- 0
14. McCarthy, M. M. (2004). Filtering the Internet: The Children's Internet Protection Act. *Educational Horizons, 82*(2), 108–113.
15. Cer, D., Diab, M., Agirre, E., López-Gazpio, I., & Specia, L. (2017). SemEval-2017 Task 1: Semantic textual similarity – Multilingual and cross-lingual focused evaluation. *arXiv: 1708.00055v1*. https://doi.org/10.48550/arXiv.1708.00055
16. Chai, C. P. (2023). Comparison of text preprocessing methods. *Natural Language Engineering, 29*(3), 509–553. https://doi.org/10.1017/S1351324922000213
17. Chandrasekaran, D., & Mago, V. (2021). Evolution of Semantic Similarity: A Survey. *ACM Computing Surveys, 54*(2), 1–37. https://doi.org/10.1145/3440755
18. Cialdini, R. B. (2021). *Influence, new and expanded: The psychology of persuasion*. Harper Business.

19. Cialdini, R. B., & Goldstein, N. J. (2004). Social influence: Compliance and conformity. *Annual Review of Psychology, 55*(1), 591–621. https://doi.org/10.1146/annurev.psych.55.090902.142015
20. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional Transformers for language understanding. In the *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL 2019, Volume 1. Long and Short Papers* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423
21. Dietterich, T.G. (2000). Ensemble methods in machine learning. In *Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science*, *1857*, 1–15. https://doi.org/10.1007/3-540-45014-9_1
22. Dshalalow, J. H. (2013). *Foundations of abstract analysis. 2nd edition*. Springer. http://dx.doi.org/10.1007/978-1-4614-5962-0
23. Directorate-General for Communications Networks, Content and Technology. (2020). *Public consultation on the AI White Paper: Final report*. European Commission. Retrieved Dec. 21, 2024, from: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68462.
24. Directorate-General for Communications Networks, Content and Technology. (2021). *Proposal for a regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act)*. European Commission. Retrieved Dec. 21, 2024, from https://digital-strategy.ec.europa.eu/en/consultations/white-paper-artificial-intelligence-european-approach-excellence-and-trust.
25. European Parliament & Council of the European Union. (2024). *Artificial Intelligence Act*. Eur-Lex. Retrieved Dec. 20, 2024, from: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32024R1689.
26. Fink, E. L., Cai, D. A., Kaplowitz, S. A., & Chung, Y. Y. H. (2003). The semantics of social influence: Threats vs. persuasion. *Communication Monographs, 70*(4), 395–421. http://dx.doi.org/10.1080/0363775032000179115
27. Floridi, L. (2002). On the intrinsic value of information objects and the infosphere. *Ethics and Information Technology, 4*(4), 287–304. https://doi.org/10.1023/A:1021342422699
28. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., & Luetge, C. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines, 28*, 689–707. https://doi.org/10.1007/s11023-018-9482-5
29. Fuller, L. L. (1964). *The morality of law* (revised ed.). Yale University Press.
30. Goitein, E. (2019). *How the FBI violated the privacy rights of tens of thousands of Americans*. Brennan Center for Justice. https://www.brennancenter.org/our-work/analysis-opinion/how-fbi-violated-privacy-rights-tens-thousands-americans
31. Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *arXiv:2301.04246*. https://doi.org/10.48550/arXiv.2301.04246

32. Gray, J. (2012). Hamann, Nietzsche, and Wittgenstein on the language of philosophers. In Anderson, L. M. (ed.), *Hamann and the tradition*, (pp. 104–121). Northwestern.
33. Gruenfeld, D. H., & Wyer, R. S. (1992). Semantics and pragmatics of social influence: How affirmations and denials affect beliefs in referent propositions. *Journal of Personality and Social Psychology, 62*(1), 38–49. https://doi.org/10.1037/0022-3514.62.1.38
34. Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. University Park Press.
35. Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques, 3rd ed*. Morgan Kaufmann.
36. Han, M., Zhang, X., Yuan, X., Jiang, J., Yun, W., & Gao, C. (2020). A survey on the techniques, applications, and performance of short text semantic similarity. *Concurrency and Computation: Practice and Experience*, *33*(5), e5971. https://doi.org/10.1002/cpe.5971
37. He, Z., Dumdumaya, C. E., Quimno, V. V. (2024). Measurement of semantic similarity. *Journal of Theoretical and Applied Information Technology, 102*(5), 1673–1685.
38. Human Rights Watch. (2019). *China's algorithms of repression: Reverse engineering a Xinjiang police mass surveillance app*. Human Rights Watch. https://www.hrw.org/report/2019/05/01/chinas-algorithms-repression/reverse-engineering-xinjiang-police-mass
39. Hursthouse, R. & Pettigrove, G. (2023). Virtue ethics, *The Stanford Encyclopedia of Philosophy, Fall 2023 Ed*. Zalta E. N. & U. Nodelman (eds.). Retrieved June 09, 2024, from https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=ethics-virtue.
40. Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., & Naaman, M. (2023). Co-writing with opinionated language models affects users' views. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI'23)*, (pp. 1–15). ACM. https://doi.org/10.1145/3544548.3581
41. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., & Liu, Q. (2020). TinyBERT: Distilling BERT for Natural Language Understanding. Findings of EMNLP 2020. *arXiv:1909.10351*. https://doi.org/10.48550/arXiv.1909.10351
42. 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
43. Jowitt, J. (2022). *Agency, morality, and law*. Hart.
44. Jurafsky, D. & Martin, J. H. (2025). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models* (3rd ed. draft). Online manuscript released January 12, 2025. https://web.stanford.edu/~jurafsky/slp3.
45. Kagan, S. (1998). *Normative ethics*. Westview. https://doi.org/10.4324/9780429498657
46. Kennedy, C. (2019). Ambiguity and vagueness. In Maienborn, C., Heusinger, K. & Portner, P. (Eds.), *Semantics: Lexical structures and adjectives*, (pp. 236–271). De Gruyter Mouton. https://doi.org/10.1515/9783110626391-008

47. Kornbeck, J. (2021). General Data Protection Regulation (GDPR) ambiguity, national diversity and data protection officer certification: Implementing Art. 39(1) GDPR in France, Italy, Luxembourg and Spain. *Journal of Data Protection & Privacy, 4* (4), 388–403.
48. Kramer, M. H. (2004). *Where law and morality meet*. Oxford.
49. Krauss, R. M., & Fussell, S. R. (1996). Social psychological models of interpersonal communication. In E. E. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 655–701). Guilford Press.
50. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A Lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*. https://doi.org/10.48550/arXiv.1909.11942
51. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*. https://doi.org/10.48550/arXiv.1907.11692
52. Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.
53. Martinich, A. P. (2016). *Philosophical writing: An introduction* (4th ed.). Wiley.
54. Meireles, A. (2022). *A brief analysis of the Brazil's data protection law*. Oxen Privacy Tech Foundation. https://optf.ngo/blog/a-brief-analysis-of-the-brazils-data-protection-law
55. Mihalcea, R., Corley, C., & Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI Proceedings of the National Conference on Artificial Intelligence, 2006, Volume 1* (pp. 775–780). AAAI.
56. Musch, S., Borrelli, M., & Kerrigan, C. (2023). The EU AI Act as global artificial intelligence regulation. SSRN. http://dx.doi.org/10.2139/ssrn.4549261
57. Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology, 19*, 123–205. https://doi.org/10.1016/S0065-2601(08)60214-2
58. Postema, G. J. (2022). *Law's rule: The nature, value, and viability of the rule of law*. Oxford
59. Qamar, U., & Raza, M. S. (2024). *Applied text mining*. Springer. https://doi.org/10.1007/978-3-031-51917-8
60. Radbruch, G. (2006). Statutory lawlessness and supra-statutory law (1946). *Oxford Journal of Legal Studies, 26*(1), 1–11. https://doi.org/10.1093/ojls/gqi041
61. Reimers, L., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084*. https://doi.org/10.48550/arXiv.1908.10084
62. Rohatyn, D. A. (1972). The language of philosophy. *Dialectica, 26*(3/4), 293–299.
63. Rozinek, O., & J. Mareš. (2021). The duality of similarity and metric spaces. *Applied Sciences, 11*, 1910. https://doi.org/10.3390/app11041910
64. Saeed, J. I. (2016). *Semantics* (4th ed.). Wiley.

65. Saint-Charles, J., & Mongeau, P. (2018). Social influence and discourse similarity networks in workgroups. *Social Networks, 52*, 228–237. https://doi.org/10.1016/j.socnet.2017.09.001
66. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*. https://doi.org/10.48550/arXiv.1910.01108
67. Sasoko, W. H., Setyanto, A., Kusrini, & Martinez-Bejar, R. (2024). Comparative study and evaluation of machine learning models for semantic textual similarity. In *Proceedings of the 2024 8th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)* (pp. 364–369). IEEE. https://doi.org/10.1109/ICITISEE63424.2024.10730053
68. Saulwick, A., & Trentelman, K. (2014). Towards a formal semantics of social influence. *Knowledge-Based Systems, 71*, 52–60. https://doi.org/10.1016/j.knosys.2014.06.022
69. Sinnott-Armstrong, W. (2023). Consequentialism, *The Stanford Encyclopedia of Philosophy, Winter 2023 Ed*. Zalta, E. N. & Uri Nodelman (eds.). Retrieved June 13, 2024, from https://plato.stanford.edu/archives/win2023/entries/consequentialism/.
70. Slote, M. (2001). *Morals from motives*. Oxford.
71. Stahl, B. C. (2012). Morality, ethics, and reflection: A categorization of normative IS research. *Journal of the Association for Information Systems, 13*(8):636–656. https://doi.org/10.17705/1jais.00304
72. Tedeschi, J. T., & Bonoma, T. V. (2017). Power and influence: An introduction. In Tedeschi, J. T. (ed.), *The social influence processes* (pp. 1–49). Routledge.
73. Van Zyl, L. (2019). *Virtue ethics: A contemporary introduction*. Routledge.
74. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *arXiv preprint arXiv:1706.03762*. https://doi.org/10.48550/arXiv.1706.03762
75. Veale, M., & Borgesius, F. Z. (2021). Demystifying the draft EU Artificial Intelligence Act. *Computer Law Review International, 22*(4) 97–112. https://doi.org/10.9785/cri-2021-220402
76. Wang, J., & Dong, Y. (2020). Measurement of text similarity: A survey. *Information, 11*(5), 421–437. https://doi.org/10.3390/info11090421
77. Wood, N. (2020). *Virtue rediscovered: Deontology, consequentialism, and virtue ethics in the contemporary moral landscape*. Lexington Books.
78. Zalta, E. N., & Nodelman, U. (Eds.). (n.d.). *Stanford Encyclopedia of Philosophy*. Stanford University. Retrieved Jan. 22, 2025, from https://plato.stanford.edu/index.html.
79. Zhao, B., Zhang, R., & Bai, K. (2024). A Fuzzy multigranularity convolutional neural network with double attention mechanisms for measuring semantic textual similarity. *IEEE Transactions on Fuzzy Systems, 32*(10), 5762–5776. https://doi.org/10.1109/TFUZZ.2024.3427801
80. Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms*. CRC Press. https://doi.org/10.1201/b12207

81. Zong, C., Xia, R., & Zhang, J. (2021). *Text data mining*. Springer. https://doi.org/10.1007/978-981-16-0100-2

Figures

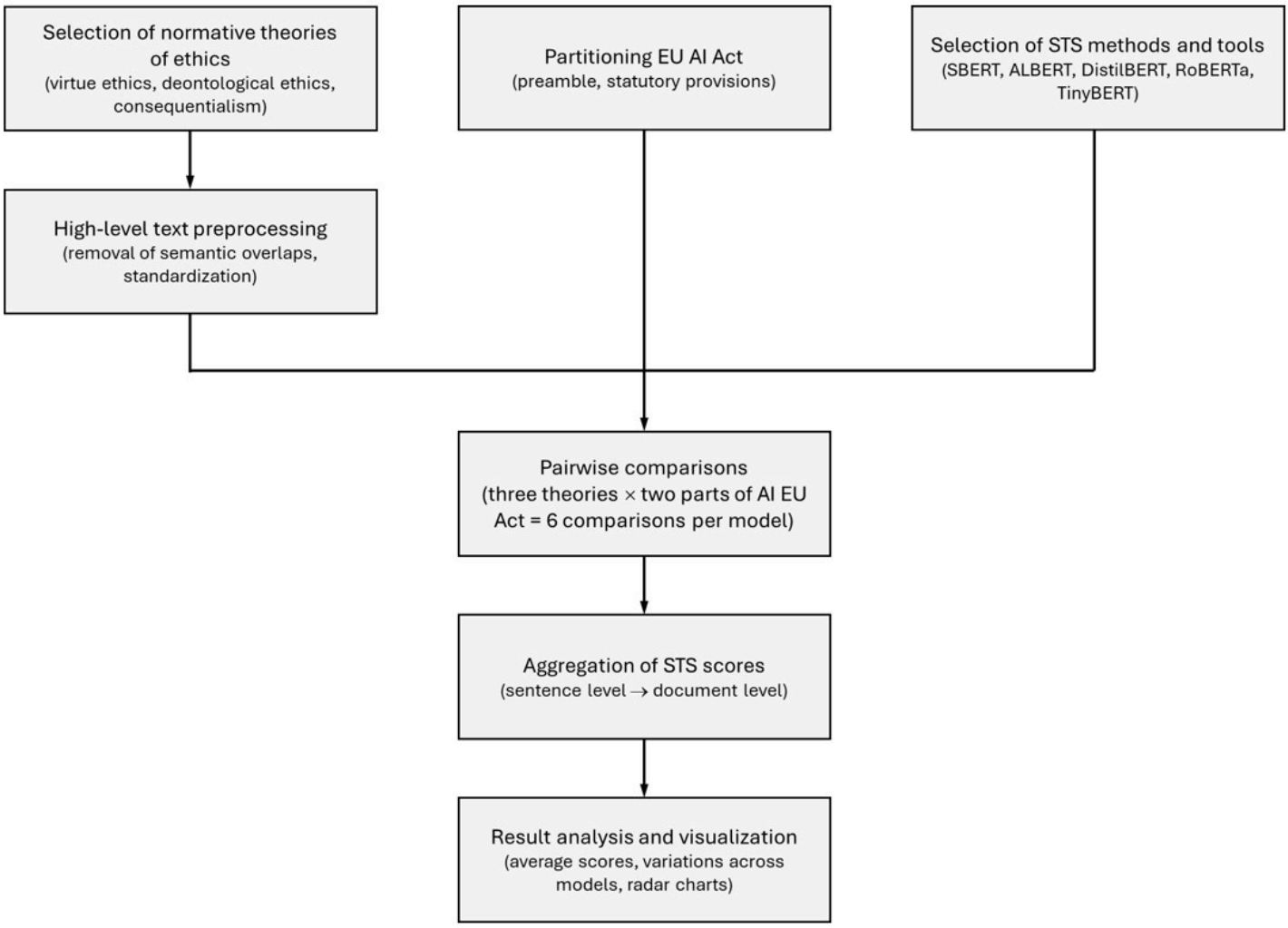
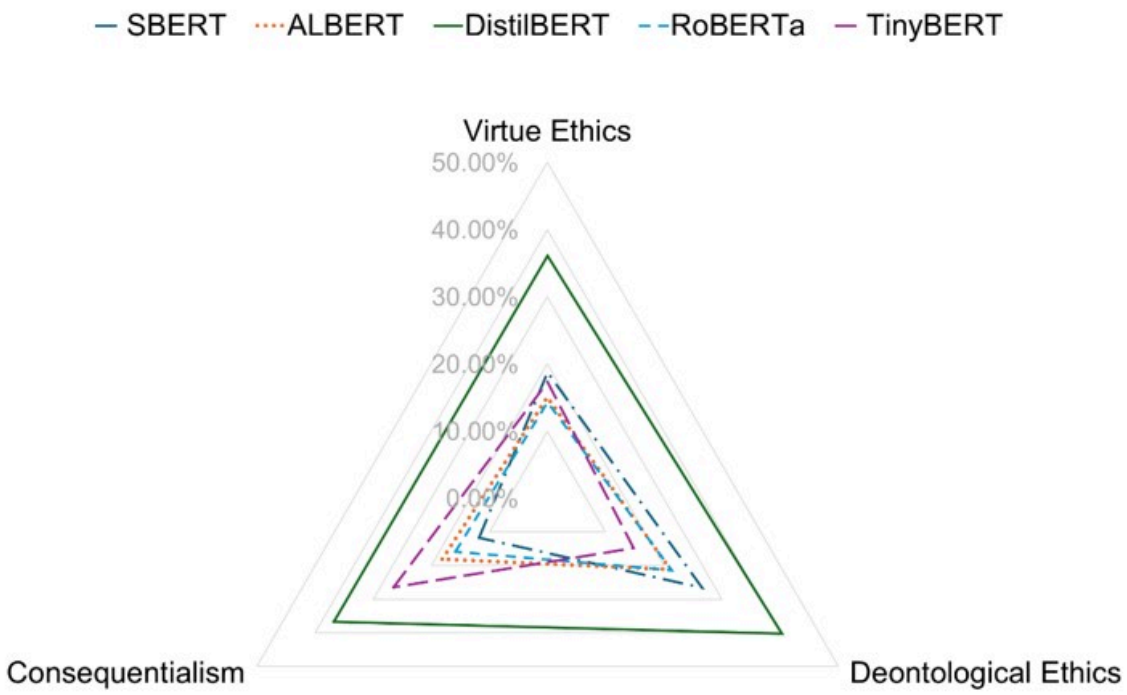
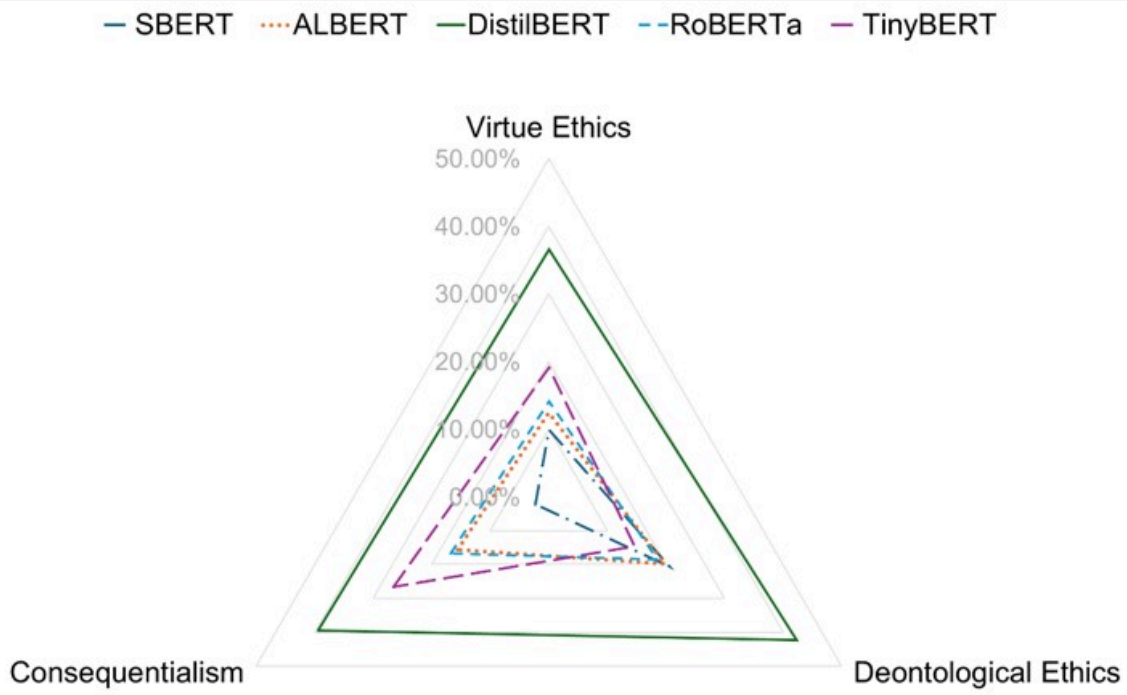


Figure 1

Flowchart of the research design and analytical workflow of the study



(a) Radar chart for the Preamble of the EU AI Act.



(b) Radar chart for the provisions of the EU AI Act.

Figure 2

Graphical summaries of the STS scores for the normative theories of ethics and **(a)** the preamble, **(b)** the provisions of the EU AI Act, respectively

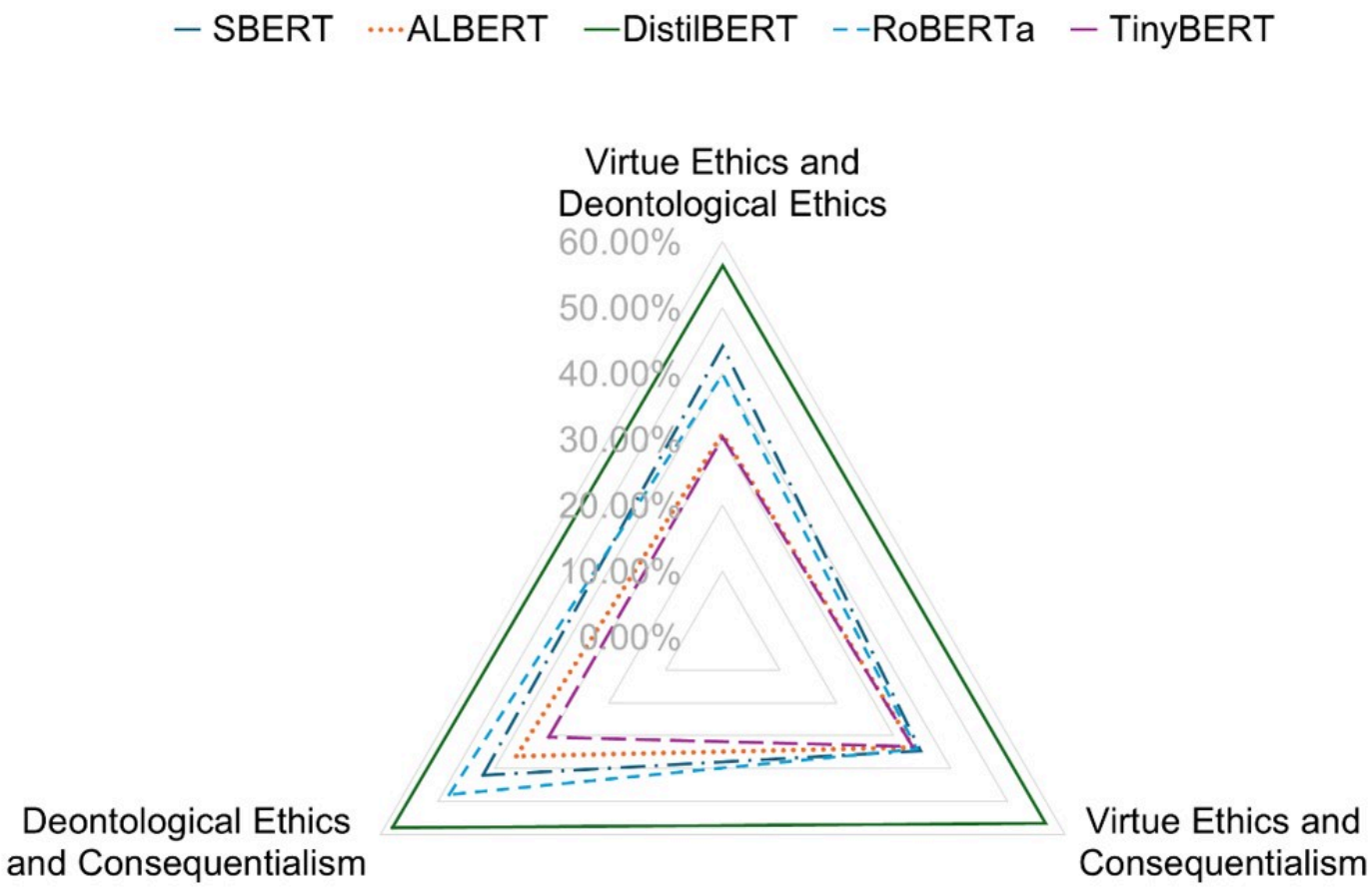


Figure 3

Graphical summary of STS scores for pairwise semantic comparisons of the three normative theories of ethics