Graph and Set Theoretic Analysis of AI Ethics Frameworks

A Comparative Study of Religious, Regulatory, Intergovernmental, and Corporate Approaches

Devina Athalia Putri Kusumah - 13525070
Program Studi Teknik Informatika
Sekolah Teknik Elektro dan Informatika
Institut Teknologi Bandung, Jalan Ganesha 10 Bandung
E-mail: devinakusumah@gmail.com , 13525070@std.stei.itb.ac.id

***Abstract*—This paper looks at six well-known AI ethics frameworks: Magnifica Humanitas, the EU AI Act, the OECD AI Principles, the UNESCO Recommendation on the Ethics of Artificial Intelligence, the OpenAI Charter, and Anthropic's Responsible Scaling Policy, and compares them using tools from discrete mathematics. Each framework is treated as a set of principles drawn from a shared vocabulary of eighteen terms, and set operations together with the Jaccard similarity coefficient are used to measure how much each pair of frameworks actually overlaps. It turns out that none of the six frameworks share a single principle in common, and the religious framework shares nothing at all with either AI company's policy. A bipartite graph connecting documents to principles is then built and projected into two simpler graphs: one linking principles that tend to be endorsed together, and one linking documents that share at least one principle. Centrality measures, a connectivity check, and the Louvain community detection algorithm are applied to these graphs. The results show that Human Dignity and Human Oversight sit at the structural center of the principle graph, that the eighteen principles split naturally into three groups (safety/process norms, rights-based norms, and solidarity-based norms), and that the document graph stays fully connected even though the religious framework barely overlaps with the corporate ones. Every coding decision is backed up with evidence in an appendix. Taken together, this suggests that the global conversation around AI ethics is only partially unified, with religious and corporate frameworks sitting on the edges of a denser regulatory core.**

Keywords—dicrete mathematics, set theory, graph theory, bipartite graph, centrality, AI ethics

I. Introduction

Over the past few years, every kind of organization that touches AI (governments, regulators, tech companies, even religious institutions) has published its own document laying out what it considers ethical AI. Skimming through a few of them, I noticed they all reach for the same handful of words: "transparency," "fairness," "safety." That made me curious. Do these documents actually mean the same thing when they use those words, or are they just borrowing each other's vocabulary while emphasizing very different things underneath?

Reading six long policy documents side by side and trying to compare them from memory is doable, but it is also subjective, two people could read the same six documents and come away with different impressions of how aligned they really are. Since I am taking IF1220 Discrete Mathematics this semester, I thought this would be a good chance to try something more systematic: represent each framework as a set of endorsed principles, then actually measure the overlap using set theory and graph theory.

For this analysis I deliberately chose documents from four very different kinds of institutions: religious, intergovernmental, regional-regulatory, and corporate. I picked six documents: Magnifica Humanitas, a 2026 statement on AI ethics issued under the Holy See; the European Union's AI Act; the OECD AI Principles; the UNESCO Recommendation on the Ethics of Artificial Intelligence; the OpenAI Charter; and Anthropic's Responsible Scaling Policy.

II. Theoretical Framework

A. Sets and Sets Operations

Let U be the universal set of every ethical principle that shows up anywhere across the six documents. For each document $Di\ (i = 1, \ldots, 6)$, $Pi \subseteq U$ is simply the set of principles that document actually endorses. With these sets defined, I used four standard operations throughout the paper:

- **Intersections $\boldsymbol{Pi \cap Pj}$:** principles shared between two frameworks.
- **Union $\boldsymbol{Pi \cup Pj}$:** the combined principle coverage of two frameworks.
- **Difference $\boldsymbol{Pi \setminus Pj}$:** principles that show up in Pi but not in Pj.
- **Generalized intersection $\boldsymbol{\cap Pi}$:** principles common to all frameworks, which I will call the "core."

B. Jaccard Similiarity

To quantify pairwise overlap between two principle sets, I used the Jaccard similarity coefficient:

$$J(Pi, Pj) = |Pi \cap Pj| / |Pi \cup Pj| \quad (1)$$

Where $J = 1$ indicated identical sets and $J = 0$ indicates disjoins sets.

C. Bipartite Graphs

A graph $G = (V, E)$ consists of a vertex set V and an edge set $E \subseteq V \times V$. A graph is bipartite if V can be partitioned into two disjoints sets $V1, V2$ such that every edge connects a vertex in $V1$ to a vertex in $V2$, with no edges within $V1$ or within $V2$.

I construct a bipartite graph $B = (D \cup P, E)$ where $D = \{D1, \dots, D6\}$ is the set of documents and $P = U$ is the set of principles, with $(Di, p) \in E$ if and only if p $\in Pi$.

D. Graph Projections

Given bipartite graph B with parts D and P, the projection onto P is the graph $GP = (P, EP)$ where two principles (p, q) are connected if at least one document endorses both of them, the edge weight is just how many documents co-endorse that pair. Projecting onto D works the same way but for documents instead, giving GD.

E. Degree and Centrality Measures

For a vertex $v \in V$, the degree centrality is:

$$CD(v) = d(v) / (|V| - 1) \quad (2)$$

The betweenness centrality of v measures how often a vertex sits on the shortest path between two other vertices.

$$CB(v) = \Sigma\, \sigma st(v) / \sigma st \quad (3)$$

Where σst is the number of shortest paths from s to t, and $\sigma st(v)$ is the number of those paths passing through v.

F. Connectivity

A graph is connected if you can get from any vertex to any other vertex by following edges. A connected component is a maximal connected subgraph. The number of connected components, $k(G)$, indicates structural fragmentation: $k(G) = 1$ implies full connectivity, while $k(G) > 1$ implies the graph has broken apart into separate, disconnected clusters.

III. Methodology

A. Documents Selection

I picked six documents using three criteria: each one had to be publicly available and citable, each had to actually spell out ethical principles rather than just technical specs, and together they had to represent genuinely different kinds of institutions.

That gave a corpus spanning four institutional types: a religious/moral-doctrinal source (Magnifica Humanitas), regional binding regulation (the EU AI Act), intergovernmental law instruments (the OECD AI Principles and the UNESCO Recommendation), and corporate self-governance documents (the OpenAI Charter and Anthropic's RSP). Comparing only regulatory documents, or only corporate ones, wouldn't tell whether AI ethics discourse is one shared project or several separate ones. A corpus of six documents was chosen to keep manual coding tractable while still permitting meaningful pairwise and graph-based comparison; 15 pairwise comparisons and a bipartite graph.

B. Coding Scheme

Each document Di was read in full and coded against a master vocabulary U of ethical principles. The vocabulary was constructed inductively: my first pass through all six documents pulled out every candidate principle term, and then I merged terms that different documents were clearly using to mean the same thing, for example, “respect for human dignity” and “human-centered values” both became Human Dignity. This produced a master vocabulary of eighteen principles: Human Dignity, Common Good, Subsidiarity, Solidarity, Social Justice, Transparency, Accountability, Human Oversight, Fairness, Safety, Privacy, Robustness, Non-discrimination, Sustainability, Inclusive Growth, Proportionality, Broad Benefit, and Cooperative Orientation.

For each document, a principle $p \in U$ was included in Pi if the document explicitly named or substantively defined that principle as a guiding value, regardless of the exact wording, as long as it mapped onto one of the eighteen terms. A principle was merely mentioned in passing, without being framed as a guiding commitment, was not included. Table I shows the resulting coding for all six documents.

This Coding scheme has an inherent limitation: judgements about whether a document “substantively” endorses a principle involve interpretation, and a different coder might produce a marginally different table. This limitation is discussed further in Section V-E. To mitigate it, the coding criteria above were applied uniformly and conservatively across all six documents, and the master vocabulary was fixed before coding began to avoid post-hoc adjustment.

C. Manual Concept Extraction Procedure

Concept extraction was performed manually, in three passes, rather than by automated keyword search, because the principles of interest are often expressed through paraphrase or through an entire clause rather than a single fixed term.

Pass 1 (Vocabulary induction): I read all six documents and underlined every passage that seemed to articulate a guiding ethical commitment. From these passages, a candidate term was extracted, either the document’s own label (e.g., the EU AI Act’s recitals explicitly name “human oversight” and “transparency” as requirements) or, if no single label existed, a short phrase summarizing what the passage was getting at (e.g., Magnifica Humanitas repeatedly invokes the protection of “human dignity”

without using that exact phrase as a section heading). This pass produced an initial list of 27 candidate terms across all six documents.

Pass 2 (Consolidation): Next, I grouped the 27 candidate terms into equivalence classes where different documents used different surface terms for what was judged to be the same underlying idea. For example, the OECD's "human-centered values and fairness" and the EU AI Act's "respect for human dignity" were both mapped to the single label Human Dignity; the OECD's "accountability" and Anthropic's commitment to external accountability mechanism were both mapped to Accountability. Each equivalence class was assigned a single canonical label, chosen to be the term most frequently used across the corpus or where no term dominated, the most semantically neutral option. This pass reduced the 27 candidates to the eighteen term master vocabulary U reported in Section III-B.

Pass 3 (Coding): with U fixed, I went back through each document a second time and, for each term in U a binary decision was recorded: present (1) if the document contained a passage mapping to that term under the Pass 2 equivalence classes, and absent (0) otherwise. A passage was only counted if it framed the term as a guiding value or requirement; for instance, a passing mention of "privacy laws" in a list of existing regulations did not count as the document endorsing Privacy as its own guiding principle, but a line saying AI systems "must respect individuals' privacy" was. The output of Pass 3 is the set of vectors $P1, \dots, P6$, summarized in Table I.

As a worked example: I coded Magnifica Humanitas as endorsing Human Oversight because it states that decisions with significant consequences for human persons should remain subject to human judgement and responsibility. It was not coded as endorsing Privacy, Transparency, or Safety, even though the document clearly cares about the ethical use of AI in general, no passage was found that framed any of these three terms specifically as a named guiding commitment in the sense defined above.

Document	Principles	$\|Pi\|$
Magnifica Humanitas (MH)	Human Dignity, Common Good, Subsidiarity, Solidarity, Social Justice, Human Oversight	6
EU AI Act (EU)	Human Dignity, Transparency, Accountability, Human Oversight, Fairness, Safety, Privacy, Robustness, Non-discrimination	9
OECD AI Principles	Human Dignity, Fairness, Transparency, Accountability, Robustness, Human Oversight, Privacy, Sustainability, Inclusive Growth	9
UNESCO Recommendation	Human Dignity, Safety, Privacy, Human Oversight, Transparency, Accountability, Fairness, Non-discrimination, Sustainability, Proportionality	10
OpenAI Charter	Safety, Broad Benefit, Cooperative Orientation, Accountability, Transparency	5
Anthropic RSP	Safety, Accountability, Robustness, Transparency, Proportionality	5

Table I. Coded Principle Sets Pi

D. Computational Pipeline

The coded sets $P1, \dots, P6$ were used to compute the following, in order:

1) Set operations: the universal set U (the union of all Pi), the core set (the intersection of all Pi), and all 15 pairwise intersections, unions, and Jaccard similarities (defined in Section II).
2) Bipartite graph construction: a graph $B = (D \cup U, E)$ with $D = \{D1, \dots, D6\}$ and $(Di, p) \in E$ if and only if $p \in Pi$.
3) Projections: GP, the projection of B onto U (principles co-endorsed by at least one document) and GD, the projection of B onto D (documents sharing at least one principle), each with edge weights equal to the number of shared principles.
4) Centrality: degree centrality and betweenness centrality computed on GP and degree centrality computed on GD.
5) Connectivity: the number of connected components of GP and GD.
6) Community detection: the Louvain algorithm applied to GP (using edge weights) to identify clusters of frequently co-endorsed principles.

All computations were performed in Python using the NetworkX library for graph construction, projection, centrality, and the python-louvain library for community detection.

E. NetworkX Implementation

This subsection lists the core NetworkX code used to implement Steps 2-6 of the pipeline above, operating, on the coded sets documents = {"MH": {…}, "EU": {…}, …} from Table I.

Bipartite graph construction (Step 2):

```
import networkx as nx
from networkx.algorithms import bipartite

U = set.union(*documents.values())
D = set(documents.keys())

B = nx.Graph()
B.add_nodes_from(D, bipartite=0)
B.add_nodes_from(U, bipartite=1)
for doc, principles in documents.items():
    for p in principles:
        B.add_edge(doc, p)B.add_nodes_from(D,
bipartite=0)
B.add_nodes_from(U, bipartite=1)
for doc, principles in documents.items():
    for p in principles:
```

Projections (Step 3):

```
# Principle co-occurrence graph
G_P = bipartite.weighted_projected_graph(B, U)

# Document similarity graph
G_D = bipartite.weighted_projected_graph(B, D)
```

Centrality (4):

```
degree = nx.degree_centrality(G_P)
betweenness = nx.betweenness_centrality(
    G_P, weight='weight')

degree_doc = nx.degree_centrality(G_D)
```

Connectivity (5):

```
nx.is_connected(G_P)
list(nx.connected_components(G_P))

nx.is_connected(G_D)
list(nx.connected_components(G_D))
```

Community detection (6):

```
from community import community_louvain

partition = community_louvain.best_partition(
    G_P, weight='weight', random_state=42)
```

IV. Results

A. Set-Theoretic Results

The universal set U contains 18 distinct principles. The core set (the intersection of all six Pi) is empty: no single principle is coded as present in all six documents. The closest principle to universal endorsement is "Transparency," present in five of six documents (all except Magnifica Humanitas), followed by "Accountability" and "Human Oversight," each present in four documents.

Figure 1 shows the Jaccard similarity $J(Pi, Pj)$ for all 15 document pairs. The highest similarity is between the EU AI Act and the UNESCO Recommendation ($J = 0.727$), followed by the OECD Principles paired with each of the EU AI Act and UNESCO ($J = 0.636$ for both), reflecting the close alignment of the three intergovernmental/regulatory frameworks. The lowest similarities are between Magnifica Humanitas and the two corporate frameworks: $J(MH, OpenAI) = 0$ and $J(MH, Anthropic) = 0$, meaning these pairs of principle sets are disjoint.

```
Jaccard Similarity Matrix J(Pi, Pj):
                    MH          EU        OECD      UNESCO      OpenAI   Anthropic
MH                   –       0.154       0.154       0.143       0.000       0.000
EU               0.154           –       0.636       0.727       0.273       0.400
OECD             0.154       0.636           –       0.583       0.167       0.273
UNESCO           0.143       0.727       0.583           –       0.250       0.364
OpenAI           0.000       0.273       0.167       0.250           –       0.429
Anthropic        0.000       0.400       0.273       0.364       0.429           –

Set differences – principles unique to each document:
  MH: {'Common Good', 'Social Justice', 'Subsidiarity', 'Solidarity'}
  EU: (none)
  OECD: {'Inclusive Growth'}
  UNESCO: (none)
  OpenAI: {'Broad Benefit', 'Cooperative Orientation'}
  Anthropic: (none)
```

Fig. 1. Jaccard Similarity $J(Pi, Pj)$. (Source: Author)

The asymmetry of Magnifica Humanitas's principle set is also evident in the set difference: of the six principles in PMH, four (Common Good, Subsidiarity, Solidarity, and Social Justice) appear in no other document, indicating that Magnifica Humanitas brings in a whole of vocabulary that nothing else in the corpus uses.

B. Bipartite Graph and Projections

The bipartite graph B has 6 (documents) + 18 (principles) = 24 vertices and 44 edges, one edge for every (document, principle) pair marked as present in Table I. and Fig. 1 show the whole thing.

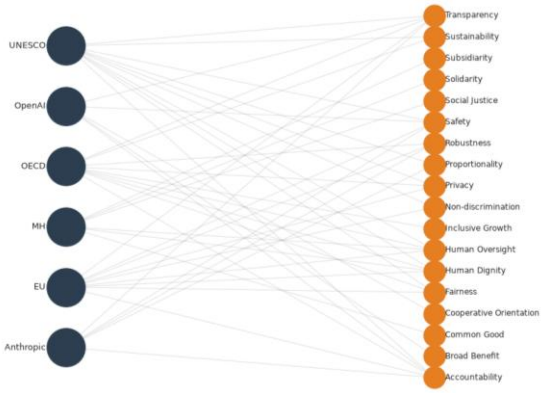

Fig. 2. Bipartite document-principle graph B. Dark blue nodes (left) are the six documents; orange nodes (right) are the eighteen principles in U. (Source: Author)

Projecting onto the 18 principles gives GP with 84 edges. Out of the 153 possible pairs of principles, 84 of them (55%) are connected by at least one document that endorses both, so the principle side of the graph is densely connected. Projecting onto the six documents gives GD with 13 out of 15 possible edges, meaning every pair of documents shares at least one principle, except for the two pairs identified in Section IV-A (J=0). Fig. 3 shows GD directly, with each edge label giving the number of principles that pair of documents has in common.

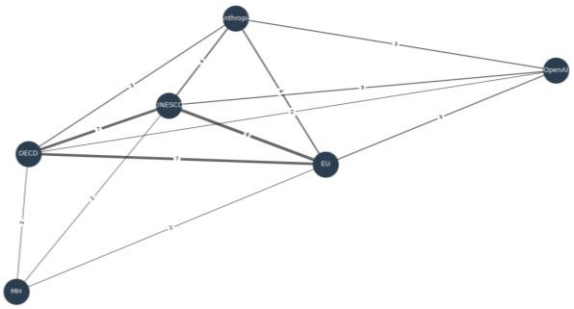

Fig. 3. Document similarity graph GD. Each document is linked to every other document with which it shares at least one principle; edge labels and edge thickness indicate how many principles each pair shares. (Source: Author)

C. Connectivity

Both projections are connected. GP has a single connected component containing all 18 principles, and GD has a single connected component containing all 6 documents. The connectivity of GD is notable given that Magnifica Humanitas shares no principle with OpenAI or Anthropic directly ($J = 0$ for both pairs). The document graph remains connected only because Magnifica Humanitas shares principles with the EU AI Act, OECD, and UNESCO, which in turn share principles with the corporate frameworks. Magnifica Humanitas is thus connected to the rest of the corpus through intermediary documents.

D. Centrality

Fig. 4 reports degree and betweenness centrality for the principle projection GP, restricted to the six highest degree principles.

```
Principle graph G_P — top 10 by degree centrality:
  Principle                 Degree CD   Betweenness CB
  ------------------------  ----------  --------------
  Human Oversight                0.882           0.178
  Human Dignity                  0.882           0.178
  Transparency                   0.765           0.058
  Accountability                 0.765           0.058
  Safety                         0.706           0.076
  Fairness                       0.647           0.001
  Sustainability                 0.647           0.034
  Privacy                        0.647           0.001
  Robustness                     0.647           0.005
  Proportionality                0.588           0.118

Principle graph G_P — top 5 by betweenness centrality:
  Principle                 Betweenness CB   Degree CD
  ------------------------  ---------------  ----------
  Inclusive Growth                   0.316       0.471
  Human Oversight                    0.178       0.882
  Human Dignity                      0.178       0.882
  Proportionality                    0.118       0.588
  Safety                             0.076       0.706

Document graph G_D — degree centrality:
  UNESCO     1.000
  EU         1.000
  OECD       1.000
  OpenAI     0.800
  Anthropic  0.800
  MH         0.600
```

Fig. 4. Centrality in GP (Source: Author)

Human Dignity and Human Oversight have both the highest degree centrality (each co-occurring with 15 of the other 17 principles) and the highest betweenness centrality in GP. By contrast, the four principles unique to Magnifica Humanitas (Common Good, Solidarity, Subsidiarity, and Social Justice) each have a degree centrality of 0.294 among the lowest in the graph.

A notable exception to the degree/betweenness correlation is Inclusive Growth, which has below median degree centrality (0.471) but the highest betweenness centrality of any principle in GP (0.316, nearly double that of Human Dignity or Human Oversight). Inclusive Growth is endorsed only by the OECD Principles, but the OECD's principle set bridges several otherwise loosely connected parts of GP, this fairly obscure principle ends up being disproportionately important for connecting everything else. This illustrates that degree and betweenness centrality can diverge: a principle endorsed by few documents can still be structurally pivotal if if those few documents are themselves well-connected.

For the document projection GD, the EU AI Act, OECD Principles, and UNESCO Recommendation each have degree centrality 1.000 (connected to all five other documents), the OpenAI Charter and Anthropic RSP each have degree centrality 0.800, and Magnifica Humanitas has the lowest degree centrality at 0.600 (connected to three of five other documents: the EU AI Act, OECD, and UNESCO, but not OpenAI or Anthropic).

E. Community Detection

Applying the Louvain algorithm to GP (using co-occurrence counts as edge weights) yields three communities, summarized in Table II.

Community	Principles
1 (Process / Safety)	Accountability, Transparency, Safety, Robustness, Proportionality, Broad Benefit, Cooperative Orientation
2 (Individual Rights)	Human Dignity, Human Oversight, Fairness, Privacy, Non-discrimination, Sustainability, Inclusive Growth
3 (Communitarian)	Common Good, Solidarity, Subsidiarity, Social Justice

Table II. Louvain Communities in GP

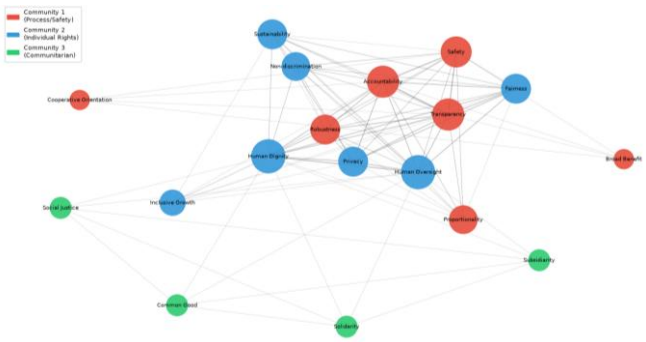

Fig. 5. Principle co-occurrence graph GP. Node size is proportional to degree centrality; color indicates Louvain community membership (communities correspond to Table II). (Source: Author)

Community 1 groups process-and safety-oriented principles, drawing heavily from the two AI-developer frameworks (OpenAI, Anthropic) along with the safety-related articles of the regulatory frameworks. Community 2 groups rights-based and individual-protection principles, drawing primarily from the EU AI Act, OECD, and UNESCO. Community 3 corresponds exactly to the four principles found only in Magnifica

Humanitas, forming a structurally distinct cluster rooted in Catholic social teaching.

V. DISCUSSION

A. An Empty Core: No universally Shared Principle

The clearest finding of this analysis is that the intersection of all six principle sets is empty: across religious, regulatory, intergovernmental, and corporate sources, no single ethical principle is coded as present in every framework. Even "Transparency," the most widely shared principle, is absent from Magnifica Humanitas. This does not necessarily indicate Magnifica Humanitas disagrees with the idea of transparency, it might just take it for granted as part of its broader commitments to human dignity and the common good, without naming it outright. However, under my coding criteria in Section III-B, it does indicate that eighteen term vocabulary used across these documents does not converge on a single common term, even among the most frequently invokes candidates.

B. Broidging Principles: Human Dignity and Human Oversight

Human Dignity and Human Oversight emerge as the structurally most central principles in GP, with the highest degree and betweenness centrality. This is significant because Human Dignity is one of only two principles (alongside Human Oversight) shared between Magnifica Humanitas and three of the other five frameworks (EU, OECD, UNESCO). In graph-theoretic terms, these two principles function as candidate cut-vertices connecting the religious framework's distinctive vocabulary (Community 3 in Table II) to the regulatory mainstream (Community 2). Their high betweenness centrality reflects this bridging role: many shortest paths between principles unique to Magnifica Humanitas and principles unique to the corporate frameworks must pass through Human Dignity or Human Oversight to get there.

C. Three Comunities: Result and Interpretation

The Louvain communities in Table II suggest a three-part structure to the ethical vocabulary of this corpus. Community 1 (Accountability, Transparency, Safety, Robustness, Proportionality, Broad Benefit, Cooperative Orientation) reflects a process-and-risk paradigm characteristic of frameworks governing the behavior of AI-developing organizations themselves. Community 2 (Human Dignity, Human Oversight, Fairness, Privacy, Non-discrimination, Sustainability, Inclusive Growth) reflects an individual-rights paradigm characteristic of regulatory and intergovernmental instruments concerned with the effects of AI on individuals and societies. Community 3 (Common Good, Solidarity, Subsidiarity, Social Justice) reflects a communitarian paradigm rooted in Catholic social teaching, emphasizing collective rather than individual goods. That these three paradigms emerge as separable communities, rather than a single undifferentiated cluster, suggests that AI ethics discourse is best understood not as one shared project but as the partial overlap of at least three distinct ethical traditions, each contributing a partially non overlapping vocabulary.

D. Connectivity Despite Disjoint Pairs

Despite $J(MH, OpenAI) = J(MH, Anthropic) = 0$, the document graph GD remains fully connected, and Magnifica Humanitas is not an isolated vertex. This illustrates a general property of graphs constructed from set overlap: pairwise disjointness does not imply global disconnection, as long as something in between connects them (intermediary sets). Here, the EU AI Act, OECD Principles, and UNESCO Recommendation each share principles with both Magnifica Humanitas and the corporate frameworks, and so function as bridges in GD in the same way that Human Dignity and Human Oversight function as bridges in GP. This suggests that intergovernmental regulatory frameworks may play a structurally mediating role between religious/ethical traditions and corporate self-governance, even where the latter two share no direct vocabulary.

E. Limitations

Three limitations of this study should be noted. First, the coding of documents into principle sets (Table I) involved interpretive judgement; while coding criteria were fixed in advance and applied uniformly, a different coder could plausibly produce a different table, particularly for borderline cases. Second, the representation here is binary: a principle is either present or absent in Pi, with no representation of the relative emphasis or priority a document assigns to a principle it endorses. A weighted representation, for example based on frequency of mention or position within the document, might reveal additional structure. Third, the corpus of six documents, while diverse, is small; a larger corpus spanning more national jurisdictions, religious traditions, and corporate actors would allow stronger claims about the generality of the three-community structure identified in Section IV-E.

VI. CONCLUSION

This paper applied set theory and graph theory to compare six AI ethics frameworks spanning religious, regulatory, intergovernmental, and corporate sources. Each framework was coded as a set of principles drawn from an eighteen-term shared vocabulary, and these sets were compared using intersection, union, and Jaccard similarity. A bipartite document-principle graph was constructed and projected into a principle co-occurrence graph and a document similarity graph, on which degree centrality, betweenness centrality, connectivity, and Louvain community detection were computed.

The results show that no principle is shared by all six frameworks, that Human Dignity and Human Oversight are the most structurally central principles in the principle graph, with Inclusive Growth playing a surprisingly important bridging role despite its low degree, that the eighteen principles decompose into three communities corresponding to process/safety, individual-rights, and communitarian paradigms, and that the document graph remains fully connected despite two pairs of documents sharing no coded principle. Together, these findings

describe global AI ethics discourse (based on this corpus and this coding) as a partially overlapping patchwork of vocabularies, held together by a handful of structurally central principles, rather than one single unified framework. Whether the resulting clusters correspond to genuinely distinct "ethical traditions" in some deeper sense is an interesting question this analysis raises

Future work could extend this analysis to a larger corpus, employ weighted or graded principle endorsement rather than a binary representation, and incorporate automated text-classification methods to reduce the subjectivity of the coding step while preserving the formal comparative apparatus developed here.

ATTACHMENT

Github Source Code of "Graph and Set Theoretic Analysis of AI Ethics Frameworks " Here

Youtube Video Here

ACKNOWLEDGMENT

The author would like to express sincere gratitude to God Almighty for His blessings, guidance, and strength throughout the process of conducting and completing this paper. Special appreciation is extended to Dr. Ir. Rinaldi Munir, M.T., lecturer of the IF1220 Discrete Mathematics course. His dedication to teaching and his expertise in discrete mathematics contributed to the development of the ideas presented in this work.

The author also wishes to thank both of the author's parents for their unwavering support, understanding, and motivation throughout the process of this work.

Finally, the author would like to acknowledge the contributions of researchers, organizations, and institutions involved in the development of AI ethics frameworks. Their efforts to establish principles for responsible and ethical artificial intelligence inspired the topic of this study and provided the basis for the comparative analysis presented in this paper.

REFERENCES

[1] Encyclical Letter "Magnifica Humanitas" of His Holiness Pope Leo XIV on Safeguarding the Human Person in the Time of Artificial Intelligence, Vatican City, 2026. https://www.vatican.va/content/leo-xiv/en/encyclicals/documents/20260515-magnifica-humanitas.html

[2] European Parliament and Council of the European Union, Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artifical Intelligence (Artificial Intelligence Act), 2024. https://artificialintelligenceact.eu/wp-content/uploads/2024/04/TA-9-2024-0138_EN.pdf

[3] Organisation for Economic Co-operation and Development, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449, 2019 (amended 2024). OECD-LEGAL-0449-en.pdf

[4] United Nations Educational, Scientific and Cultural Organization, Recommendation on the Ethics of Artificial Intelligence, 2021. https://unesdoc.unesco.org/ark:/48223/pf0000381137/PDF/381137eng.pdf.multi

[5] OpenAI, OpenAI Charter, 2018.

[6] Anthropic, Anthropic's Responsible Scaling Policy, 2023 (updated 2024). https://www-cdn.anthropic.com/616dee633636e5bd309cb73aed8622e80fe47839.pdf

[7] Munir, R. (2024). IF1220 Matematika Diskrit- Semester II Tahun 2025/2026: Graf (Bagian 1) (Versi Update 2025). https://informatika.stei.itb.ac.id/~rinaldi.munir/Matdis/2025-2026/20-Graf-Bagian1-2026.pdf

[8] R. Diestel, Graph Theory, 5th ed. Berlin: Springer, 2017. https://daiwz.net/course/disc_math/2023/Diestel_Graph_Theory.pdf

[9] K. H. Rosen, Discrete Mathematics and Its Applications, 8th ed. New York: McGraw-Hill, 2019. https://faculty.ksu.edu.sa/sites/default/files/[Book]%20Discrete%20mathematics%20and%20its%20applications%20(2019)_0.pdf

[10] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communitites in larger networks," J. Stat. Mech., vol. 2008, no. 10, P10008, 2008. https://arxiv.org/pdf/0803.0476

APPENDIX: CODING EVIDENCE

For each of the 44 (document, principle) pairs I coded as present in Table I, here's a quick note on the passage that justified that call, organized by document, so the mapping from source text to Table I can be checked independently. Section/article references point to the documents listed under References.

Magnifica Humanitas:

- **Human Dignity**—Ch. 2 frames the human person as image of God, bearer of inviolable rights and intrinsic dignity; this is the document's central anthropological premise.
- **Common Good**—Ch. 2 lists the common good as the first of the principles of Social Doctrine applied throughout the encyclical.
- **Subsidiarity**—Ch. 2 describes subsidiarity as valuing cooperation between generations, peoples, disciplines and cultures, applied to AI governance in Ch. 3.
- **Solidarity**—Ch. 2 pairs solidarity with subsidiarity as a guiding principle; Ch. 1 frames mutual care and solidarity as conditions for human flourishing.
- **Social Justice**—Ch. 2 lists social justice among the foundational principles of Social Doctrine, alongside the universal destination of goods.
- **Human Oversight**—Ch. 3 and Ch. 4 discuss AI governance and responsibility, framing decisions affecting persons as requiring human judgment and accountability.

EU AI Act:

- **Human Dignity**—Recitals affirm that AI systems must respect fundamental rights, including human dignity, as a cross-cutting requirement.
- **Transparency**—Title III/Art. 13 imposes transparency obligations on high-risk AI systems (information to users, traceability of outputs).

- **Accountability**—Title III establishes obligations for providers/deployers, including conformity assessments and record-keeping, operationalizing accountability.
- **Human Oversight**—Art. 14 requires high-risk AI systems to be designed to allow effective human oversight during use.
- **Fairness**—Recitals and Art. 10 address bias and discrimination in training data as part of the Act's fairness requirements.
- **Safety**—Title III sets risk-management and safety requirements for high-risk AI systems as a core regulatory category.
- **Privacy**—Recitals affirm the Act applies without prejudice to GDPR and data-protection rights, and Art. 10 addresses data governance.
- **Robustness**—Art. 15 requires high-risk AI systems to achieve appropriate accuracy, robustness, and cybersecurity.
- **Non-discrimination**—Recitals and Art. 10 explicitly name discrimination and bias mitigation among the Act's objectives.

OECD AI Principles:

- **Human Dignity**—Principle 1.1 ("Inclusive growth, sustainable development and well-being") and 1.2 ("Human-centered values and fairness") frame human dignity and rights as foundational.
- **Fairness**—Principle 1.2 names fairness explicitly alongside human-centered values as a core value-based principle.
- **Transparency**—Principle 1.3 ("Transparency and explainability") is one of the five value-based principles.
- **Accountability**—Principle 1.5 ("Accountability") establishes that AI actors should be accounfi for the proper functioning of AI systems.
- **Robustness**—Principle 1.4 ("Robustness, security and safety") covers robustness as a named value-based principle.
- **Human Oversight**—Principle 1.4 and accompanying commentary discuss human oversight mechanisms as part of robustness/safety.
- **Privacy**—Principle 1.2 commentary explicitly references privacy and data protection as part of human-centered values.
- **Sustainability**—Principle 1.1 names sustainable development explicitly as part of inclusive growth.
- **Inclusive Growth**—Principle 1.1 is titled "Inclusive growth, sustainable development and well-being", naming inclusive growth directly.

UNESCO Recommendation:

- **Human Dignity**—Sec. III ("Human dignity and autonomy") is one of the named values underpinning the Recommendation.
- **Safety**—Sec. III ("Safety and security") is listed among the core values.
- **Privacy**—Sec. III ("Right to privacy, and data protection") is a named value.
- **Human Oversight**—Sec. IV policy area on "Governance" calls for human oversight and the ability to attribute responsibility for AI decisions.
- **Transparency**—Sec. III ("Transparency and explainability") is a named value.
- **Accountability**—Sec. III ("Responsibility and accountability") is a named value.
- **Fairness**—Sec. III ("Non-discrimination") and §B ("Fairness and non-discrimination" framing) underpin fairness commitments.
- **Non-discrimination**—Sec. III explicitly names non-discrimination as a core value of the Recommendation.
- **Sustainability**—Sec. III ("Sustainability") is listed as one of the ten core values.
- **Proportionality**—Sec. III ("Proportionality and Do No Harm") is the first listed value, framing proportionality as foundational.

OpenAI Charter:

- **Safety**—The Charter states that OpenAI will research the riskiest aspects of advanced AI and prioritize safety research above maximizing profit.
- **Broad Benefit**—The Charter's primary fiduciary duty is described as ensuring AGI benefits all of humanity, not OpenAI shareholders.
- **Cooperative Orientation**—The Charter commits to stop competing with and assist value-aligned, safety-conscious projects in late-stage AGI development.
- **Accountability**—The Charter commits to providing public updates on safety and policy efforts as a form of external accountability.
- **Transparency**—The Charter commits to publishing safety, policy, and standards research broadly, framed as a transparency commitment.

Anthropic RSP:

- **Safety**—The RSP's core mechanism is a tiered AI Safety Level (ASL) framework defining safety requirements scaled to model capability.

- **Accountability—**The RSP commits to external accountability mechanisms, including a Responsible Scaling Officer role and board oversight.
- **Robustness—**ASL-based requirements include red-teaming, security controls, and evaluations to ensure robustness against misuse.
- **Transparency—**The RSP commits to publishing model evaluation results and capability assessments.
- **Proportionality—**The ASL framework explicitly scales safety/security requirements proportionally to assessed model risk.

PERNYATAAN

Hereby I declare that this paper that I have written is my own work, not a reproduction or translation of someone else's work and not plagiarized.

Bandung, 1 Juni 2025

Devina Athalia Putri Kusumah
13525070