Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A hyper-knowledge graph system for research on AI ethics cases

Chuan Chen [*], Yu Feng, Mengyi Wei, Zihan Liu, Peng Luo, Shengkai Wang, Liqiu Meng

Chair of Cartography and Visual Analytics, Technical University of Munich, Munich, Germany

ARTICLE INFO

Keywords:
AI ethics
Hyper-knowledge graph system
Case-oriented ontological model
Knowledge-based system
Visual analytics

ABSTRACT

Current studies on the artificial intelligence (AI) ethics focus either on very broad guidelines or on a very special domain. Therefore, the research outcome can hardly be converted into actionable measures or transferred to other domains. Potential correlations between various cases of AI ethics at different granularity levels are unexplored. To overcome these deficiencies, the authors designed a case-oriented ontological model (COOM) and a hyper-knowledge graph system (HKGS) for the research of collected AI ethics cases. COOM describes criteria for modelling cases by attributes from three perspectives: event attributes, relational attributes, and positional attributes on the value chain. Based on it, HKGS stores the correlation between cases as knowledge and allows advanced visual analysis. The correlations between cases and their dynamic changes on value chain can be observed and explored. In HKGS's implementation part, one of the collected ethics cases is used as an example to demonstrate how to generate a hyper-knowledge graph and to visually analyze it. The authors also anticipated how different practitioners of AI ethics, can achieve the desired outputs from HKGS in their diverse scenarios.

1. Introduction

1.1. Motivation

AI technology is increasingly integrated into human life. Products and services based on AI technology are playing an essential role in human society. Some exhibit anthropomorphic features as a substitute for manual work. Others demonstrate unique capabilities that exceed those of humans, such as massive computing at a high speed. So far, ethics has served as a fundamental framework for the functioning of human society. Now, this framework needs to contemplate the role of AI products and services that are increasingly embedded in human society. In fact, this demand is prevalent in various domains. AI ethics issues have already arisen to a large extent in the domains of intelligent service robots, language or vision models and autonomous driving [1], requiring more intensive investigations and practical solutions. Existing related work is lacking in providing engineering contributions to the study of AI ethics through the perspective of the knowledge-based system.

[*] Corresponding author.
E-mail addresses: chuan.chen@tum.de (C. Chen), y.feng@tum.de (Y. Feng), mengyi.wei@tum.de (M. Wei), khan.liu@tum.de (Z. Liu), peng.luo@tum.de (P. Luo), shengkai.wang@tum.de (S. Wang), liqiu.meng@tum.de (L. Meng).

https://doi.org/10.1016/j.heliyon.2024.e29048
Received 17 January 2024; Received in revised form 21 March 2024; Accepted 28 March 2024
Available online 3 April 2024

1.2. Related work

The systematic study of AI ethics issue necessitates a strict definition and the establishment of core values to minimize discrepancies in the understanding of certain concepts, such as transparency [2] and positive or negative effect [3]. Segun analyzed the thematic distinction among robot ethics, machine ethics and computational ethics in a rigorous manner [4]. The understanding of AI ethics issues can also be influenced by different philosophical, cultural, and historical backgrounds [5]. No wonder, the determination of core values in AI ethics issues has been controversial. The most widely accepted suggestion is that AI ethics values should be aligned with the universal values of human society by involving more people with sufficient diversity in this research [6]. Much of the analysis of AI ethics issues results in the creation and/or extension of guidelines [7]. These guidelines are a very good blend of core values [8,9]. However, the guidelines are often too general and hierarchically all-inclusive to be actionable [10–14].

Existing researches on AI ethics are mainly concentrated on a specific domain corresponding to individual ethical values and concerns. For instance, Arnold et al. and Aïvodji et al. investigated the fairness of AI algorithms with biases and proposed improvements [15,16]. The biases can be related to gender, race and disabilities [17–19]. Themes such as privacy protection [20] and autonomous driving safety have caught much attention as well.

Another trend of research on the ethics issues of AI is reflected in the creation and refinements of general guidelines, which converge largely [8]. Many researchers found the ambiguity in the translation of these guidelines into practice [9]. The difficulty in implementation stems from the fact that guidelines are not specific enough [12]. As an illustrative example, the term "Explainability" is clear at the macro level when mentioned in the guidelines. However, for researchers in various domains who are concerned with the measures required to achieve "Explainability" in their respective industries, the use of the term in the guideline lacks specificity and a seamless transition from macro to micro.

Further research works are reported in related domains around AI technology. Some researchers suggested to explore the AI ethics in relationship with various business practices [21]. Other researchers appealed for studies on cross-stakeholder AI ethics cases [22]. The visualization of legal results is also considered as a means of implementation [23,24].

Fig. 1 illustrates the current research gap, with a distinct focus on macro-level ethical guidelines. These guidelines primarily address overarching cases and large clusters, lacking the necessary granularity when examining individual or domain-specific cases. Conversely, research on domain-specific issues tends to produce detailed findings and conclusions, but these findings often struggle to generalize, limiting their broader impact. The research gaps revealed in existing studies have triggered the development of COOM and HKGS. Unlike the general guidelines, the COOM focuses on attributes describing individual cases on a tiny scale. The HKGS treats a specific case as the basic unit of study, which is not isolated, but linked with other units in a knowledge graph. HKGS also heuristically provides suggestions to users by relating cases with each other based on their attributes. It may facilitate a holistic understanding of a case. In this way, HKGS realizes the focus on the intricate mapping connection in AI ethics, connecting the macro-level and the micro-level.

AI ethics cases in the COOM are described by attributes from three perspectives: event attributes, relational attributes and positional attributes on the value chain. The detailed explanation of these attributes will be provided in the subsequent chapter. The analysis of event attributes leads to the insight into the temporal and geographical distribution of cases. Relational attributes support the discovery of potential correlations among different cases through link prediction based on graph convolutional network (GCN) [25]. Positional attributes describe the data flow at different processing stages of AI approaches, corresponding to value change in business conducts. This allows for the cross-domain analysis of AI ethics issues. With these features coupled with existing methods,

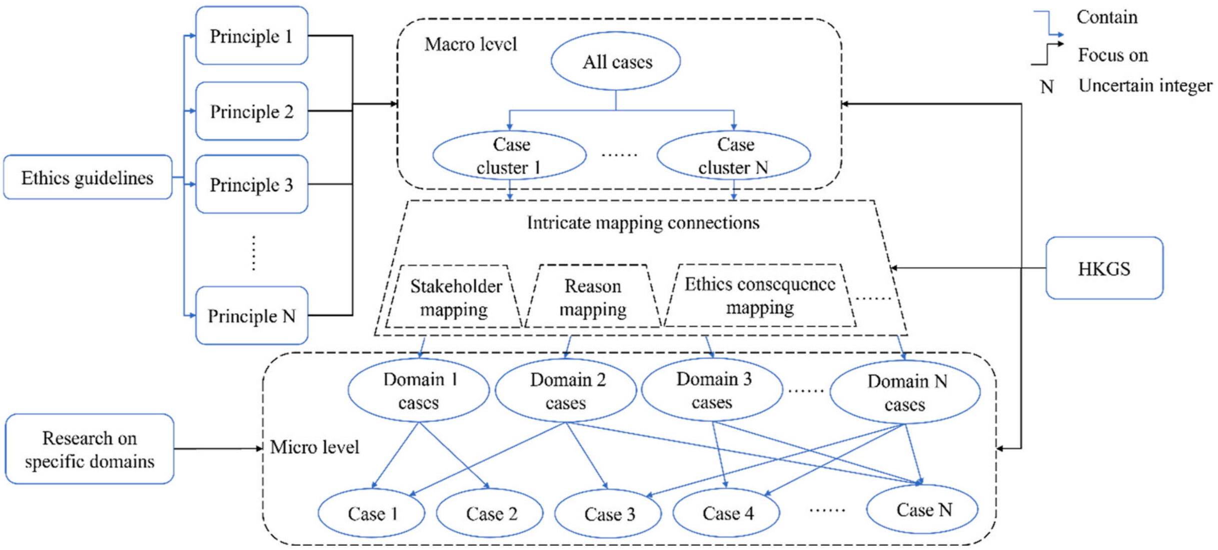


Fig. 1. HKGS focuses on a coherent macro to micro perspective in AI ethics.

HKGS can provide a coherent view on AI ethics as required by researchers [26]. To be more precise, HKGS will be utilized to showcase its ability to generate valuable output in various scenarios. These scenarios encompass legislation, industry, and company, where practitioners will accord greater priority to AI ethics research [27].

Moreover, visualization techniques are regarded as efficient tools that can facilitate the understanding of AI ethics. HKGS has powerful visualization capability in the output stage. The usage of positional attributes allows HKGS to display the evolution of AI ethics cases along the value chain. Each ethics case appears as a node, while the edges connecting cases represent correlations. Correlations based on different attributes in the HKGS can be visualized, giving users an intuitive overall impression of AI ethics cases.

2. Methodology

2.1. Experimental dataset

In total, 148 AI ethics cases are extracted from an AI incident database [28]. AI incident database is an open source and extensible dataset. The purpose of the database is to build an analysis system that includes incidents to ensure the use of AI. At the time of the experiment, there were 156 existing cases. Each case contains only a limited number of objective attributes and the description of the case. So the dataset used in this research is a new dataset based on it, which will be open source. The database in this research, consisting of 148 cases, was carefully curated through manual review, eliminating those unrelated to AI ethics. A criterion is applied during the case review process. Specifically, an AI ethics case must involve interactions between AI systems and humans that result in the loss or protection of human emotions, bodies, rights, or property. A content analysis-based annotation method was used for each attribute in each case. Multiple AI ethics researchers labeled the attributes of the cases while calculating Krippendorff's alpha to determine the reliability of the labeling results [29]. Detailed descriptions of the case attributes and Krippendorff's alpha values are presented in subsequent content.

2.2. Case-oriented ontological model

The case-oriented ontological model (COOM) serves to model AI ethics cases. It provides criteria for the hierarchical and logical relationships between case attributes. Each case is described by attributes covering three perspectives - event attributes, relational attributes and positional attributes in the value chain as shown in Fig. 2.

Based on the aforementioned contents, a coherent viewpoint from macro to micro will be constructed. Such a continuous perspective requires correlations between the cases to be analyzed. Such a coherent viewpoint requires correlations between the cases to be analyzed. The reference frames needed for the correlation analysis include spatio-temporal space, feature space, and the value chain axis, which are corresponded to these three perspectives - event attributes, relational attributes and positional attributes in the value chain.

Cases are analyzed and studied in spatio-temporal space based on event attributes. Relational attributes refer to those attributes that can be algorithmically encoded and converted into tensors for analysis in feature space. Positional attributes in the value chain are special because they can be encoded in feature space and, more importantly, they can be used to study cases along the value chain.

Event attributes refer to the basic characteristics of ethics cases as events. They include the index of a case in the database, the

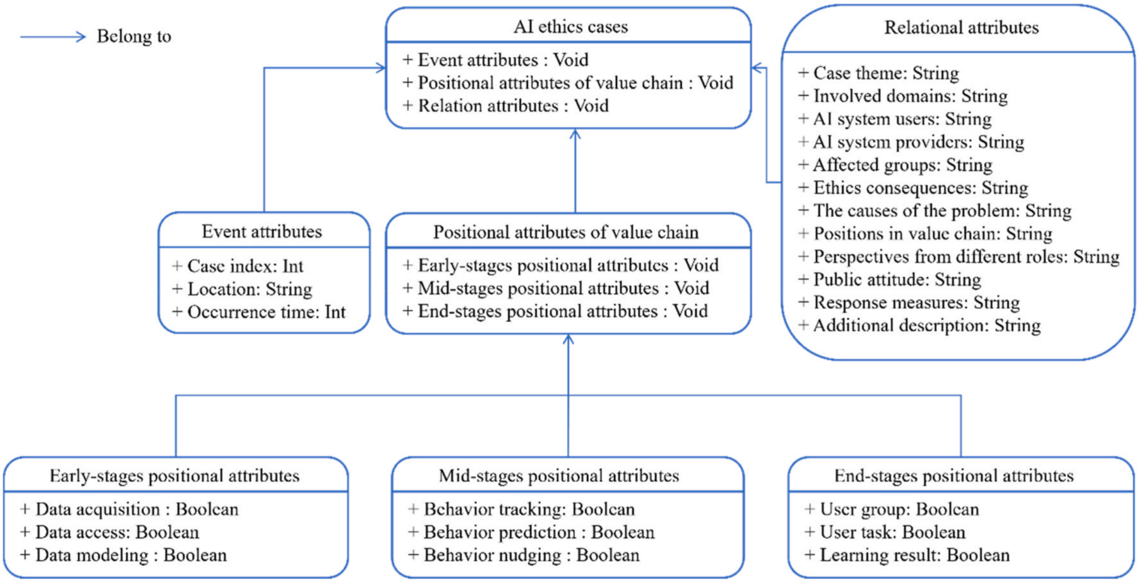

Belong to
AI ethics cases
+ Event attributes : Void
+ Positional attributes of value chain : Void
+ Relation attributes : Void
Relational attributes
+ Case theme: String
+ Involved domains: String
+ AI system users: String
+ AI system providers: String
+ Affected groups: String
+ Ethics consequences: String
+ The causes of the problem: String
+ Positions in value chain: String
+ Perspectives from different roles: String
+ Public attitude: String
+ Response measures: String
+ Additional description: String
Event attributes
+ Case index: Int
+ Location: String
+ Occurrence time: Int
Positional attributes of value chain
+ Early-stages positional attributes : Void
+ Mid-stages positional attributes : Void
+ End-stages positional attributes : Void
Early-stages positional attributes
+ Data acquisition : Boolean
+ Data access: Boolean
+ Data modeling : Boolean
Mid-stages positional attributes
+ Behavior tracking: Boolean
+ Behavior prediction : Boolean
+ Behavior nudging : Boolean
End-stages positional attributes
+ User group: Boolean
+ User task: Boolean
+ Learning result: Boolean

Fig. 2. It shows the structure of COOM in a UML diagram.

location, and the time of occurrence. Event attributes are objectively observable. Because they are independent of the perceptions of parties involved, such as alleged perpetrators, victims, and other stakeholders, they can be used directly for statistical analysis. The detailed definition of event attributes is shown in Table 1.

Relational attributes describe the semantic meanings of a case through relations. They allows for comparative analysis between cases. The attribute "case theme" relates a case with a main subject. It is usually derived from the headline or synopsis of the case story. For example, the topic of case102 [30] in our collected database is "Tesla's Full Self-Driving tech keeps getting fooled by the moon, billboards, and Burger King signs". "Involved domains" relates a case with impacted domains such as autonomous driving, AI supervision, intelligent recommendation. "AI system users and providers" relates a case with the character roles. "Affected groups" relates a case with the type of influenced population. The causes and consequences of ethics problem are collected in the report of each case. Views from different roles, public attitudes and response measures are reflected in the comments and reactions from all parts of society. Case descriptions other than the above attributes are categorized as the additional description section. The detailed definition of relational attributes is shown in Table 2.

Along the value chain, the value embedded in the data grows as the stages of data processing advance. In an early stage, the value is relatively low as the raw data may be noisy, flawed and messy. This early stage is concerned with data acquisition, data access and data modeling. When it comes to the mid stage, the data have gained certain business value and application significance, often based on the understanding of observed human behavior, including behavior tracking, behavior prediction and behavior nudging. In the final stage, the data reach the maximum value and can directly help the commercial product designer create products and services for right user groups and user tasks. AI systems can be iteratively enhanced to generate greater value. In the iterative learning process, some AI products may spontaneously learn negative content and lead to surprising learning results. For example, an AI chatbot ends up being a racist or terrorist as the learning result from a corresponding corpus. Positional attributes in the value chain describe the location of a case in the data processing sequence. The detailed definition of positional attributes in the value chain is shown in Table 3.

Event attributes are individually labeled due to the objectivity. According to content analysis theory, labeling data is judged to be fully reliable when its Krippendorff's alpha is higher than 0.800 while this data is judged to be suitable for drawing preliminary conclusions when its Krippendorff's alpha is only higher than 0.667. Krippendorff's alpha of relational and positional attributes averages 0.8697 and 0.8962. The lowest Krippendorff's alpha of them is 0.8013, higher than 0.800. The confidence level of the attributes in this dataset is higher than the useable standard, which is shown in Fig. 3.

Case 102 is taken as an example to show how the cases are structured with COOM in Table 4.

2.3. Construction of hyper-knowledge graph system (HKGS)

Based on COOM, each case is modeled by the attributes as shown in Fig. 2. Aforementioned attributes are described in a textual form. Event attributes are used for time-based and location-based statistics. Relational attributes and positional attributes of value chain are used to generate the hyper-knowledge graph system (HKGS).

Based on bidirectional encoder representations from transformers (BERT) [31], the relational attributes of a case are converted into a tensor representation. BERT has several characteristics that will be more appropriate for processing text data.

(1) Transformer Architecture: BERT is based on the Transformer model, a deep learning model that introduces Self-Attention, which enables global modeling of input sequences, making the model more effective in dealing with long-range dependencies.
(2) Bidirectional context modeling: Unlike traditional nature language processing (NLP) models, BERT employs bidirectional context modeling. It learns word representations in context through a pre-training phase, which allows each word to be represented taking into account its left and right contexts, helping to better understand ambiguities and complex relationships in language.
(3) Contextual representation: The contextual representation generated by BERT contains rich semantic information, so it is able to capture more complex semantic relationships in many NLP tasks, improving the generalization ability of the model.

In this research, BERT is used in following steps as shown in Fig. 4.

Suppose there is a sentence S containing n words or tokens, then it can be expressed as:

$$S = [w_1, w_2, w_3, \ldots, w_n]$$

where w_i denotes the i th word or token in the sentence.

BERT converts each word or token w_i in the sentence into a corresponding word tensor or token tensor x_i through a word embedding layer:

Table 1
The detailed definition of event attributes.

Attributes	Definitions
Case index	Index of case retrievals in database
Location	Location where the case occurred
Occurrence time	Time when the case occurred

Table 2
The detailed definition of relational attributes.

Attributes	Definitions
Case theme	The main subject of the case, usually rewritten in the actual label based on the case news headline
Involved domains	The domains involved in the case. The attributes do not limit the number of domains.
AI system users	The direct users of the AI system in the case
AI system providers	The direct providers of the AI system in the case
Affected groups	The actual population affected in the case
Ethics consequences	The implications of the ethics issues in the case due to the AI system
The causes of the problem	The direct cause of the ethics issue in the case
Positions in the value chain	Stages on the value chain the case occurs on
Perspectives from different roles	Views about the case from different roles in the community
Public attitude	The public attitude about the case in the community
Response measures	The case's response measures to the ethics issue
Additional description of the case	Additional information other than the above attributes will be included in it

Table 3
The detailed definition of positional attributes in value chain.

Attributes	Definitions
Data acquisition	The process of collecting, acquiring, or entering data into a system.
Data access	The ability to retrieve and obtain data from storage or data sources.
Data modeling	The process of describing and representing data through the creation of models.
Behavior tracking	The monitoring and recording of user or system activities, interactions, and behaviors.
Behavior prediction	The prediction of future user or system behavior by analyzing existing behavioral data.
Behavior nudging	The design of a system's interface, feedback, or other mechanisms to motivate a user or system to perform specific behaviors in a desired direction.
User group	The collection of target users of a system who may have similar characteristics, needs, or behaviors.
User task	A specific operation or activity that a user performs in a system.
Learning result	The knowledge or patterns that an AI system extracts from data through learning algorithms.

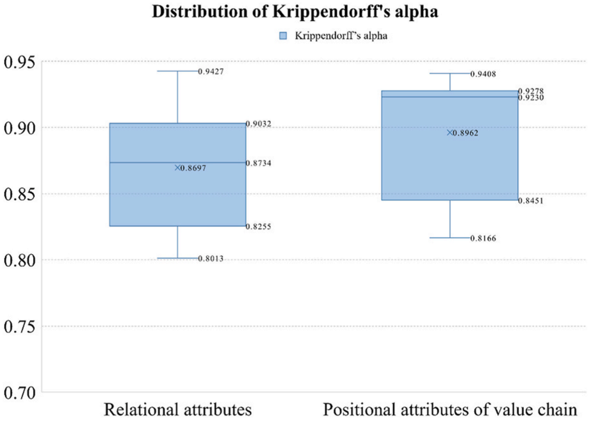


Fig. 3. It represents the distribution of Krippendorff's alpha of relational attributes and positional attributes of value chain.

$$x_i = Embedding(w_i)$$

The BERT model encodes the input sequence through multiple Transformer layers. Each Transformer layer performs self-attention mechanism and feed-forward neural network operations. The word tensors for each word or token are weighted, combined and transformed to give a new encoded tensor e_i.

$$e_i = TransformerLayer(x_i)$$

For the whole sentence S, a representation tensor E consisting of all encoded tensors e_i can be obtained.

$$E = [e_1, e_2, e_3, \ldots, e_n]$$

For a single case, the i th relational attribute is converted to the case tensor E_i. The whole case is converted to the tensor C.

$$C = [E_1, E_2, E_3, \ldots, E_{12}]$$

Table 4
The structure of Case 102 implementing COOM.

Attribute Types	Attributes	Content
Event attributes	Case index	102
	Location	America
	Occurrence time	2021, 07, 23
Relational attributes	Case theme	Tesla's Full Self-Driving Tech keeps getting fooled by the moon, billboards, and Burger King signs
	Involved domains	Autonomous Driving
	AI system users	Tesla drivers
	AI system providers	Tesla
	Affected groups	Anyone involved in Tesla vehicles and the transportation environment around Tesla
	Ethics consequences	Tesla's autopilot feature has made certain errors that have resulted in inconvenient traffic conditions for both the user and the surrounding traffic environment.
	The causes of the problem	Tesla's fully self-driving service hasn't lived up to the hype.
	Positions in the value chain	Data acquisition, Data modeling, Behavior tracking
	Perspectives from different roles	Owners: Owners have also reported their cars mistaking the sun for a red light. And one bizarre clip shows a Tesla being fooled by a truck with a traffic light. Tesla: "Full self-driving" is a $10,000 driver-assist feature offered by Tesla. While all new Teslas are capable of using the "full self-driving" software, buyers must opt for the costly add-on if they want access to the feature. The software is still in beta and is currently only available to a select group of Tesla owners, although CEO Elon Musk has claimed that a wider rollout is imminent. Musk promises that "full self-driving" will be fully capable of getting a car to its destination in the near future.
	Public attitude	Negative and confusing
	Response measures	Tesla: Further improvements will be made in the follow-up to perfect this fully self-driving service.
	Additional description of the case	Tesla's fully autonomous driving technology has had several problems in city tests. The system hesitated at intersections, tried to drive against traffic, and required frequent manual intervention. Despite Tesla's claims that it will change the world, real-world experience shows that it performs poorly in complex city traffic. In tests, the vehicle nearly collided with a construction site, attempted to ram a parked truck, and even experienced an emergency braking maneuver. Despite having a human driver on standby, the fully automated system often needs to be manually shut down to avoid dangerous situations. Tesla's fully autonomous driving software is still in beta testing and only available to a small number of Tesla owners. The technology has been described as impressive but flawed, sometimes performing well and sometimes performing dangerously. In tests in the urban environment of New York City, the system struggled to adapt to complex road conditions, hesitating and braking harshly, causing dissatisfaction among surrounding vehicles. Fully automated driving technology still needs to be improved in the city to be comparable to human drivers. The experience was unsettling and highlighted the challenges of driving vehicles with AI in subtle details.
Positional attributes in value chain	Data acquisition	Existence
	Data access	Non-existence
	Data modeling	Existence
	Behavior tracking	Existence
	Behavior prediction	Non-existence
	Behavior nudging	Non-existence
	User group	Non-existence
	User task	Non-existence
	Learning result	Non-existence

In this article, the correlation between two cases is represented by the angle between their corresponding tensors, which is calculated by cosine similarity, since the numerical value of case tensors cannot be interpreted.

By calculating the correlation $Correlation_{m,n}$ of C_m and C_n of the m th and the n th cases in 148 cases, the node of each case is connected to the nodes of several most similar cases. In equation, $\|C_m\|_2$ denotes the L2 paradigm of C_m.

$$Correlation_{m,n} = (C_m \bullet C_n) \,/\, (\|C_m\|_2 \times \|C_n\|_2) \quad 1 \leq m, n \leq 148$$

This leads to an initial graph as the input of the framework. Nodes in the initial graph embody the relational attributes and the edges represent the apparent correlations based on the cosine similarity of case tensors.

Further workflow starting from initial graph is shown in Fig. 5.

The link prediction task is completed by a two-layer Graph Convolutional Network (GCN). During the construction of GCN, the negative sampling is implemented, which slices the dataset into a positive part and a negative part [32]. After semi-supervised learning training, a trained GCN model can be generated. The training was repeated ten times and the corresponding metrics are shown in Fig. 6. Training accuracy averages 93.00%, while validation accuracy averages 92.42% and test accuracy averages 91.69%. This performance has been maintained over multiple training sessions.

The representation tensor E_i is the input of GCN encoder part and the output is labeled $E_i^{'}$. $C^{'}$ is combined from $E_i^{'}$.

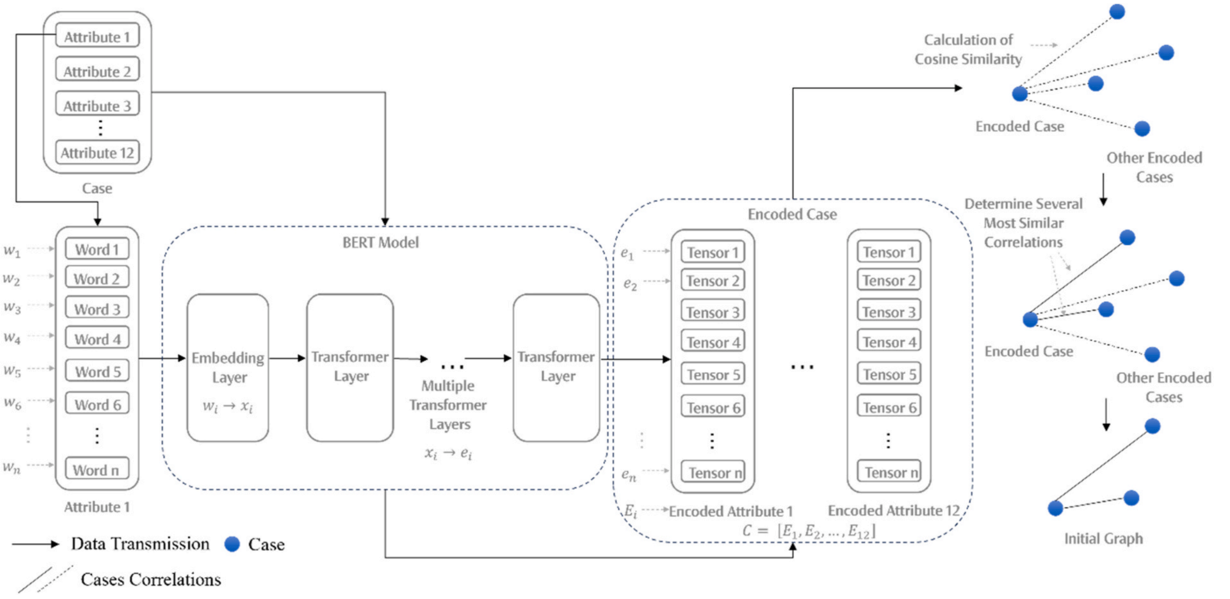


Fig. 4. It shows the construction of initial graph by BERT model.

$$E_i^{'} = GCNLayer(E_i)$$

$$C^{'} = [E_1^{'}, E_2^{'}, E_3^{'}, \ldots, E_{12}^{'}]$$

The potential correlations are mined through the decoding part. For one case pair Case *m* and Case *n* where the link prediction results show they having potential correlations, the correlation of each relational attributes $RC_i^{m,n}$ can be expressed as:

$$RC_i^{m,n} = \left(E_i^{m^{'}} \bullet E_i^{n^{'}}\right) / \left(\left\|E_i^{m^{'}}\right\|_2 \times \left\|E_i^{n^{'}}\right\|_2\right) \quad 1 \leq m, n \leq 148; \; 1 \leq i \leq 12$$

Where $E_i^{m^{'}}$ and $E_i^{n^{'}}$ represent feature tensors of the *i* th relational attributes of Case *m* and Case *n* after encoded by GCN. Therefore, the output of the link prediction is a graph containing potential correlations generated by relational attributes.

Using positional attributes, AI ethics case nodes are arranged along the value chain axis. As shown in Fig. 3, the red, green and yellow square frame represent the early stage, mid stage and end stage respectively. Relational features extracted by GCN, correlations with other nodes and value chain information along the positional attributes axis are integrated into each node. The graph that contains nodes with such properties is called hyper-knowledge graph. It allows visual analysis along the axis of positional attributes or within the plane of event attributes and relational attributes.

With the plane of event attributes and relational attributes, correlations in the aforementioned knowledge graph can be deciphered in a bar chart. Inside the chart, correlations are visualized for individual relational attributes. The general correlation between two nodes can thus be interpreted intuitively.

3. Results

3.1. Basic statistical results of AI ethics cases

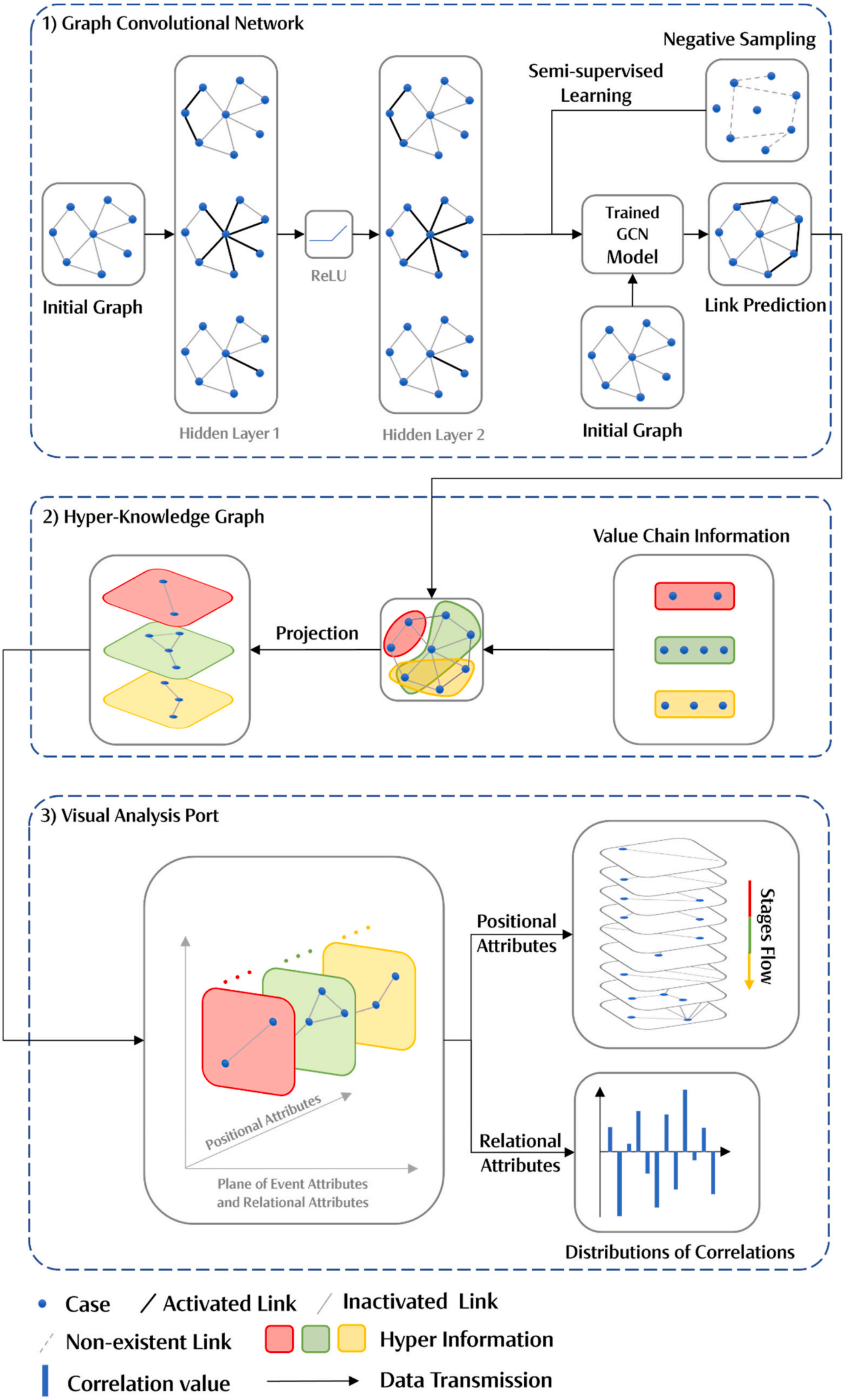
3.1.1. Statistical results of AI ethics cases by time of occurrence

The occurrence time and the locations of AI ethics cases in dataset are analyzed statistically. In Fig. 7, AI ethics cases are categorized at intervals in the time span in which AI has flourished.

Fig. 4 shows that our collected AI ethics cases grow exponentially over time. Less than 5 percent of AI ethics cases were reported before 2013, when the prevalence of AI in human life was relatively limited, and thus public awareness and media sensitivity to AI ethics issues were relatively low. Cases that occurred between 2013 and 2015 constituted over 10 percent of the dataset, with the number tripling between 2016 and 2018. Cases occurring between 2019 and 2022 account for more than half of the dataset, which can be attributed to the proliferation of media reports and the growing integration of AI products into human lives.

3.1.2. Statistical results of AI ethics cases by locations

Fig. 8 illustrates the level of concern about AI ethics issues by geographic location. It is noteworthy that a larger number of cases



(caption on next page)

Fig. 5. The workflow of HKGS starting from initial graph.

from a particular region does not suggest that more AI ethics disasters arise in that region. It is not possible either to estimate statistically how many cases are unreported or undiscovered. The available data only provides insights into the varying levels of concern regarding AI ethics issues across different regions and the corresponding sensitivity of the media in those regions towards this matter. Some cases in dataset do not have a well-defined place of occurrence. For instance, cases happening on the Internet may simultaneously affect people in numerous regions. These cases are assigned to the 'Global' category.

The percentage of cases in the dataset that occurred in the United States reaches 42.07%, indicating the highest concern and media sensitivity to AI ethics issues there. The second majority of the cases is from 'Global', which accounts for 23.78% of the whole dataset. A substantial proportion of cases are reported in China and the United Kingdom. This fact may indicate a tendency in these regions to focus on the ethics issues of AI. The remaining locations which have reported cases confirm a rather broad geographic distribution of AI ethics issues. Developed countries and regions also tend to see more coverage and social discussion of AI ethics issues.

3.1.3. Statistical results of AI ethics cases by positional attributes along the value chain

The distribution of positional attributes of AI ethics cases in the value chain are demonstrated in Fig. 9. It shows that AI ethics cases occur in early stages more frequently with the highest percentage of 92.57% in the data modeling stage. The percentages of cases in data acquisition and data access stages are 77.03% and 37.84%. Approximately one-tenth of AI ethics cases occur in the tracking,

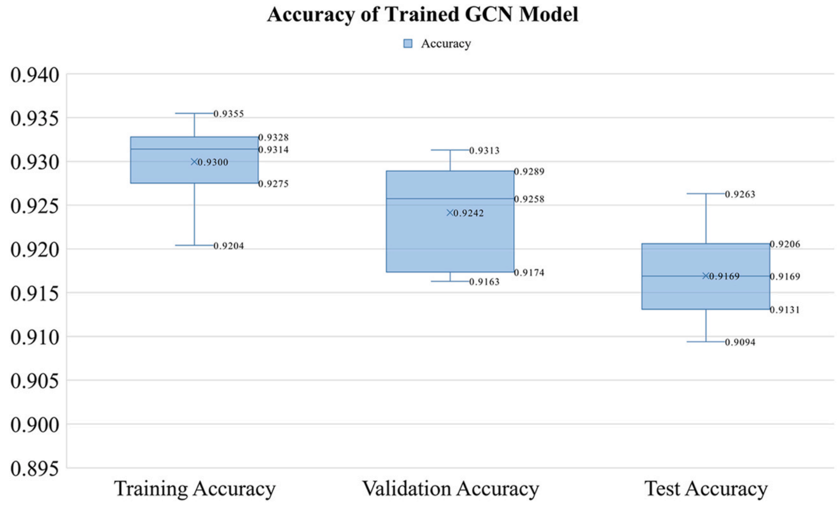


Fig. 6. It shows the accuracy distribution of training set, validation set and test set.

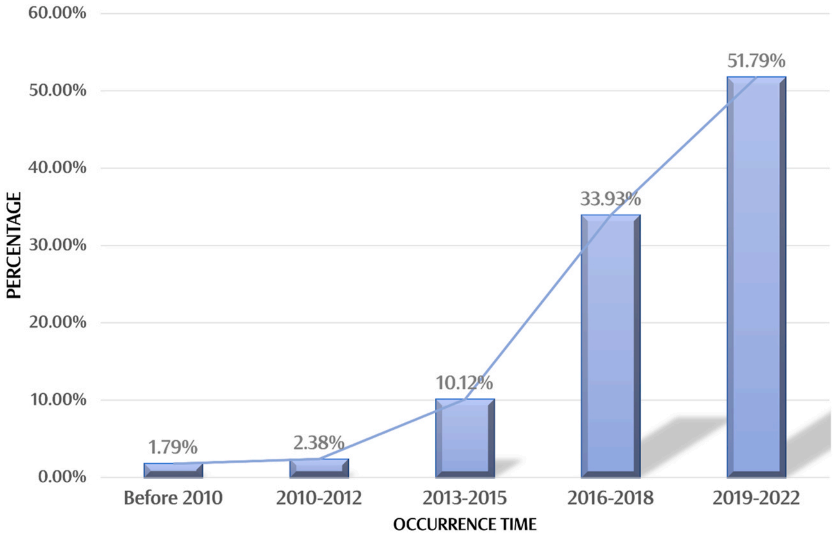


Fig. 7. The distribution of collected AI ethics cases by time of occurrence.

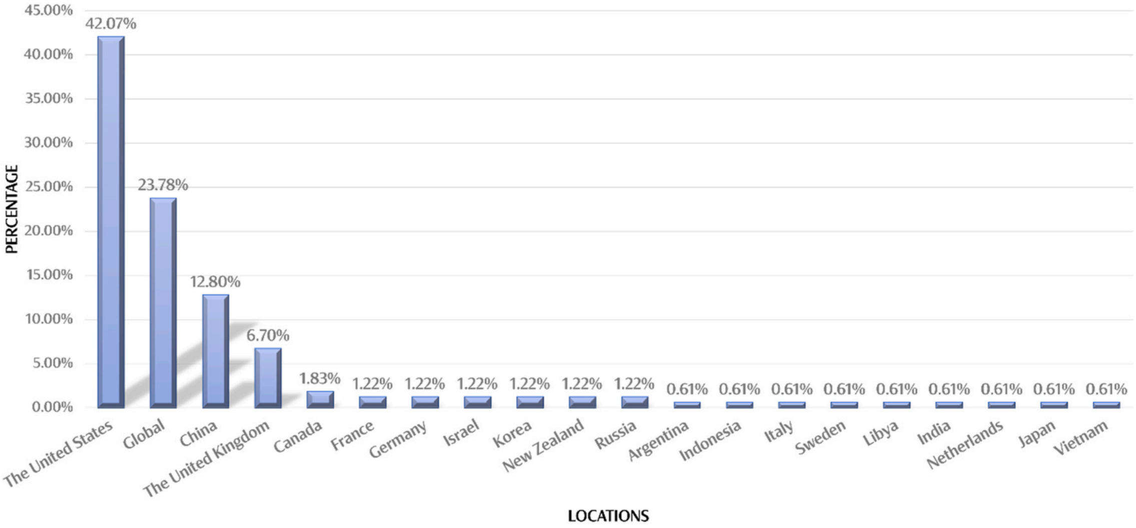


Fig. 8. The distribution of collected AI ethics cases by locations.

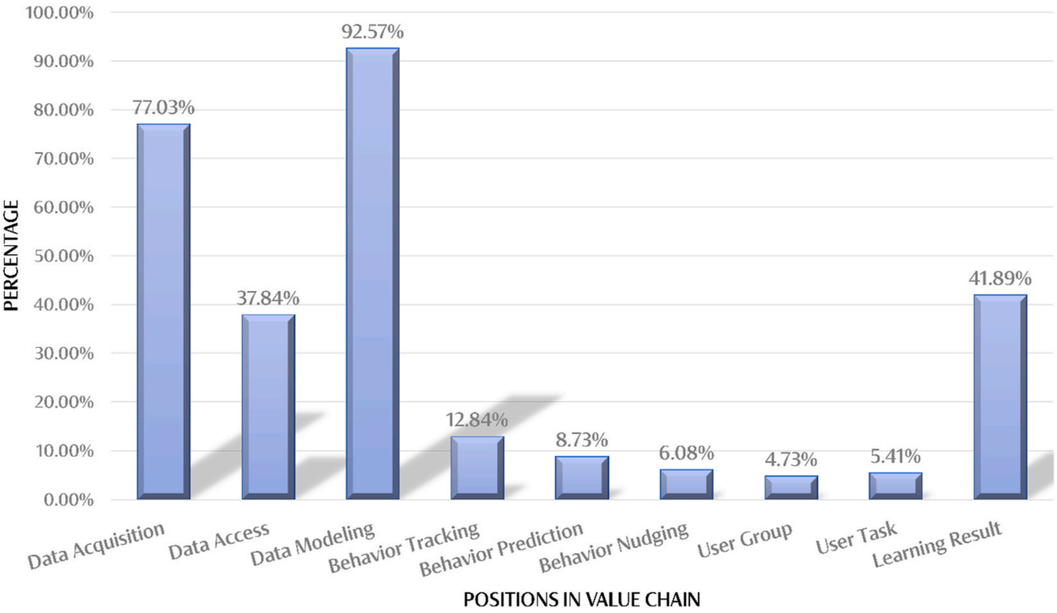


Fig. 9. The distribution of AI ethics cases by positional attributes along the value chain.

prediction and nudging of behavior. About five percent of cases occur due to the misunderstanding in the user's group and task. Noteworthy is that 41.89% of the cases are accompanied by surprising results of learning algorithms. It suggests that AI can give rise to unforeseen consequences during the learning process, and the impact of such an AI ethics case is rather high.

3.2. HKGS-based analysis of a single case

Case 102 is taken as an example for visual analysis. It tells the story of how Tesla's Autopilot mistook the moon for a yellow stoplight, reflecting the immaturity of the full autopilot that could not yet match the product description. The entire hyper knowledge graph is shown in Fig. 10.

By projecting the case onto nine stages along the value chain, we get nine panels of knowledge graph as depicted in Fig. 11.

In each panel, nodes represent cases that are most closely related to Case 102 and the relative strength of correlation is represented by the darkness of the edges. Case 102 exhibits potential strong correlations with other cases in eight of nine stages along the value chain. Only the panel corresponding to the stage "user group" is empty, indicating no strong correlations with other cases. In Fig. 11,

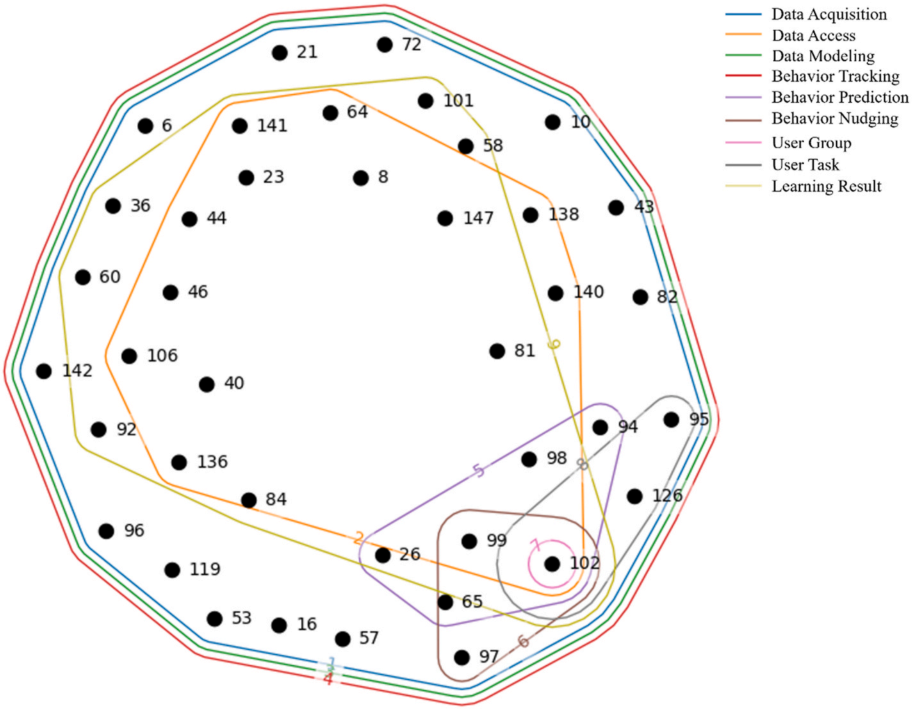


Fig. 10. The hyper knowledge graph of Case 102.

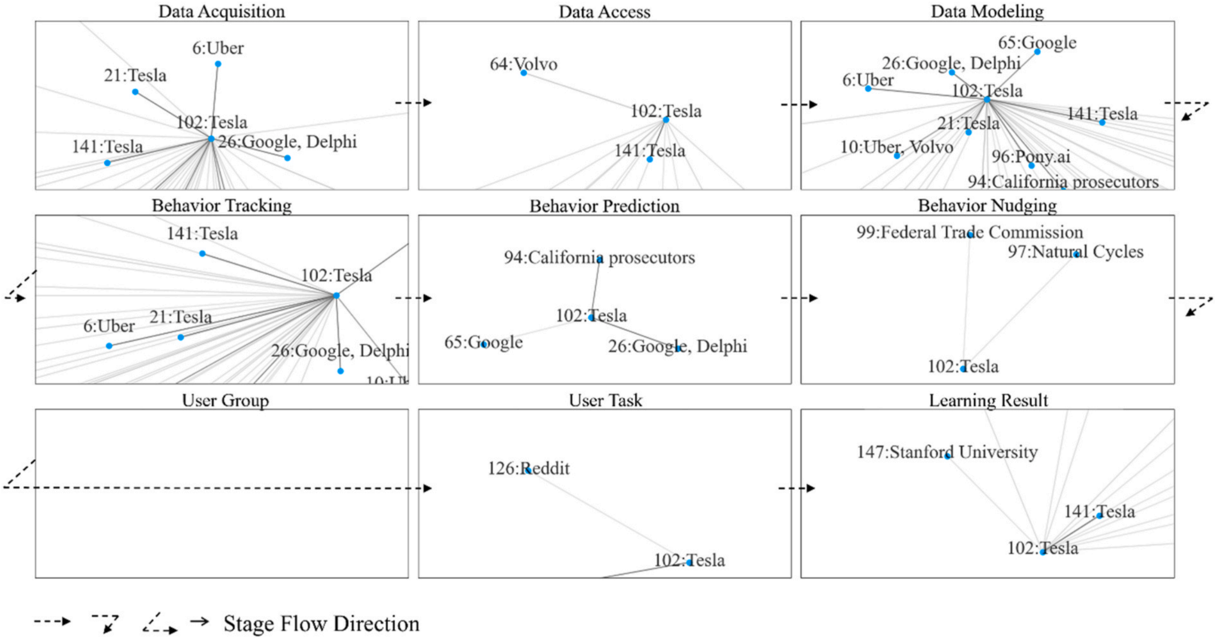


Fig. 11. The output of Case 102 by HKGS.

taking the stage "data access" as an example, it can be seen that Case 102 has potential strong correlations with Case 141 and Case 64, and the correlation with Case 141 is stronger than with Case 64. Case 21 has the same stakeholders and is correlated with Case 102 in stages "data acquisition", "data modeling" and "behavior tracking" [33].

Further information about correlations between individual relational attributes of Case 102 and Case 21 is computed and made accessible, e.g. using a bar chart as shown in Fig. 12. Both cases revolve around themes related to Tesla, not surprisingly, their major elements have to do with the drawbacks of autonomous driving. They demonstrate a strong correlation in "AI system provider",

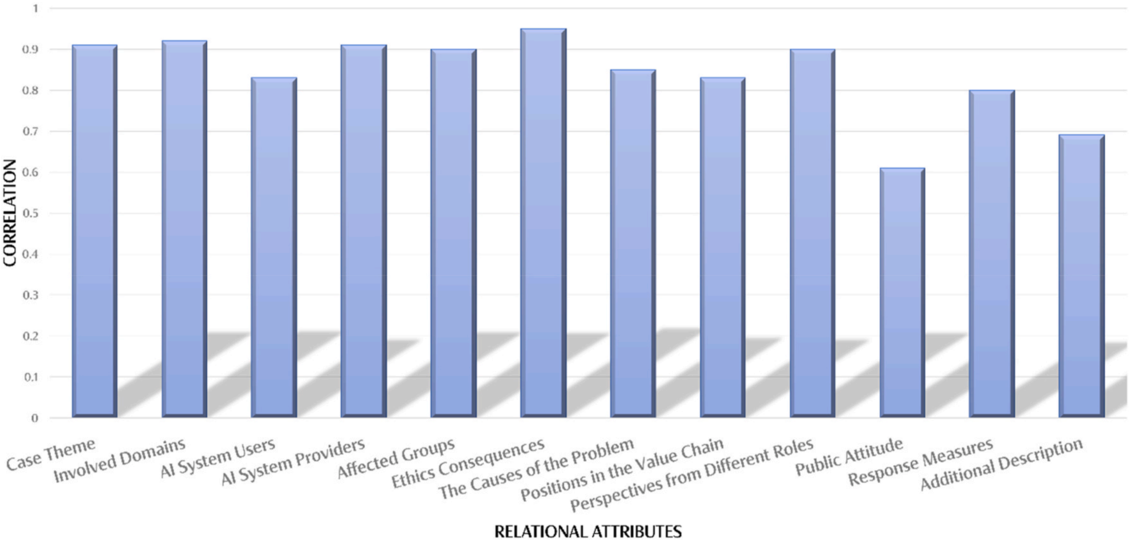


Fig. 12. The correlations between relational attributes of Case 102 and Case 21.

"affected groups", "ethics consequences" and "perspectives from different roles". The most affected people in both cases are the drivers and passengers on the vehicles. The reports of the two cases have also revealed that both cases have an impact on human safety and convenience while different roles believe that the technology of autonomous driving needs to improve. This interpretation based on case reports can be verified by the high correlations in the corresponding attributes shown in Fig. 12.

Another representative example is Case 140, which exhibits potentially strong correlations with Case 102 across multiple stages of the value chain, including data acquisition, data access, data modeling and behavior tracking [34].

Fig. 13 displays the correlations between relational attributes of Case 102 and Case 140. The correlations regarding case theme, domains are weak. Since the two cases have different themes and stakeholders, their correlations regarding AI system users, providers and affected groups are also weak. However, the HKGS reveals strong correlations between the two cases in ethics consequences, the cause of the problem, positions in value chain and perspectives from different roles. Case 140 describes chatbots that appear to make racist comments in conversations with people after learning some negative human verbiage. The ethics consequence is that the technical limitations of AI affect the experience of the product and cause annoyance to human life, which is strongly correlated to ethics consequence of Case 102. Both cases reflect the inadequate learning ability of AI from data, such as street view images and human language. Both cases describe the shortcomings of a commercial product, so the cases occur in a similar position in the value chain. The prevailing opinions on two cases not only affirm the original intent of AI product features, but also expect that AI technology may become more mature in the future. On these relational attributes, the results of the HKGS are consistent with the media opinions.

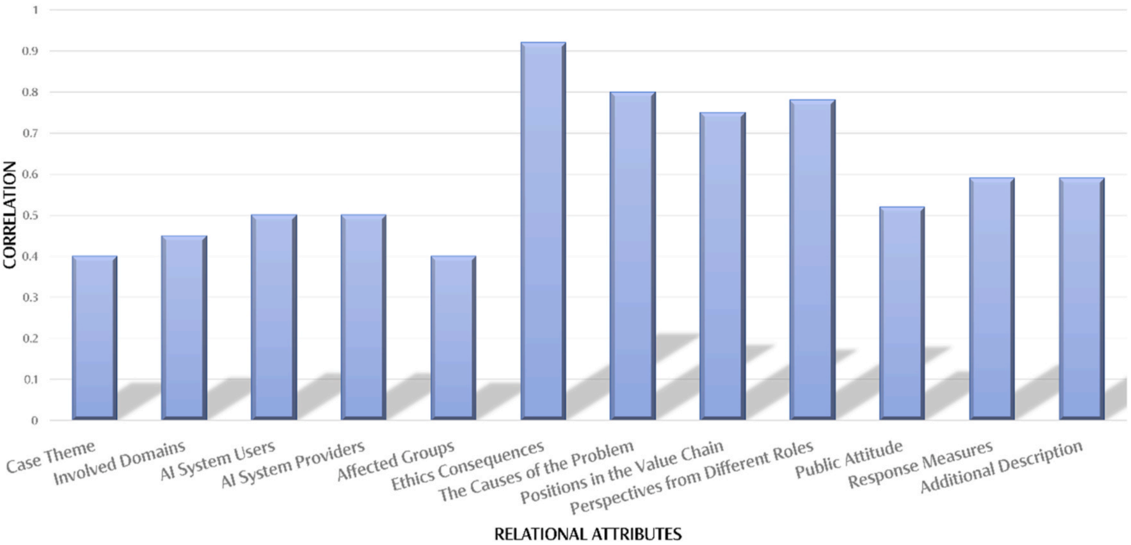


Fig. 13. The correlations between relational attributes of Case 102 and Case 140.

3.3. Different scenarios analysis

To validate the performance of HKGS, the authors chose three scenarios to demonstrate how practitioners in these scenarios can utilize HKGS to acquire meaningful results as shown in Fig. 14. In the scenario of legislation, practitioners can explore the interrelation of cases across various domains. The results will assist practitioners in developing comprehensive norms or regulations. In the scenario of industry, practitioners can analyze the correlations of cases in a domain. As a consequence, they can summarize the norms of AI ethics issues within a specific domain. These outputs can contribute to the development of the domain-specific regulation of AI ethics. In the scenario of company, practitioners can analyze the correlations between cases within a company. These outputs will aid the company in regulating their business practices and mitigating potential AI ethics risks. In conclusion, practitioners in different scenarios can derive meaningful analytical capabilities from HKGS.

3.3.1. Industry scenario: HKGS-based analysis of cases within a specific domain

In this section we take a close look at a group of cases in a certain domain, with the focus on the internal correlation within the group.

3.3.1.1. Case study: autonomous driving. Autonomous driving is selected as an example to demonstrate our analysis. For the *i* th relational attributes, all case pairs Case *m* and Case *n* within autonomous driving domains where the link prediction results show they having potential correlations are chosen. The correlations of relational attributes $RC_i^{m,n}$ of all these case pairs are averaged to $\overline{RC_i}$:

$$\overline{RC}_i = \sum RC_i^{m,n} \Big/ N$$

Where *N* represents the number of these case pairs like Case *m* and Case *n*.

The general distribution of correlations between different autonomous driving cases are displayed in Fig. 15.

These cases exhibit strong correlations regarding AI system users, AI system providers, and affected groups. One of the reasons is the scarcity of manufacturers equipped with self-driving capabilities as well as a relatively homogeneous group of victims in such cases. Besides these, the position of these cases on the value chain is also strongly correlated. The attitude of the public towards these cases is very similar. To prove this point from another perspective, Fig. 16 shows their distribution of positional attributes of value chain. It can be seen that in the early stages, the ethics issues of autonomous driving mainly occur in the data acquisition and data modeling stages. In the middle stage, these cases are highly accompanied by behavior prediction. 15.79% of the cases occur with surprising learning results.

Fig. 17 shows another output of HKGS in value chain. The cases of autonomous driving mainly focus on data acquisition, data modeling and behavior prediction, which corresponds to the distribution in Fig. 17. However, in the stages of data access, behavior tracking, behavior nudging, and user group, there are several strong correlations due to the limited number of available cases.

3.3.2. Legislation scenario: HKGS-based analysis of cases across domains

This section is dedicated to the analysis of case clusters from different domains.

Firstly, seven popular domains are selected to conduct a general multi-domain case analysis [1]. Fifty percent of all AI ethics issues are attributed to occurrences in these seven domains. They include autonomous driving, intelligent recommendation, smartphone, AI recruitment, AI supervision, face recognition and predictive policing, as shown in Fig. 18 with its abbreviation in brackets. The color intensity indicates the strength of correlations. It can be seen that cases of autonomous driving, intelligent recommendation, AI recruitment and predictive policing have a stronger internal correlation compared to other domains. The aforementioned correlation between cases of autonomous driving and smartphone is relatively high in Fig. 18. Similar to this, cases of predictive policing and AI supervision also have a strong correlation. Besides, the correlations between cases of AI recruitment and smartphone, AI recruitment and AI supervision, face recognition and smartphone are relatively strong.

The correlation of cases in different domains is usually weak. However, strong correlations may exist in some special domain pairs as mentioned above. For practitioners in the legislation scenario, this result offers information about the general correlation of cases across different domains. Since the enactment of a statute or guidance will encompass all domains, the results from HKGS will aid in formulating them macroscopically.

Then several domain pairs are chosen to display another output using HKGS. Fig. 19 shows the general correlation between autonomous driving and intelligent recommendations. These two case clusters differ greatly in AI system provider. Consequently, It can be concluded that there is hardly any overlap between the AI system providers in these two domains. Apart from it, there is not strong correlation with any other attributes. Such a result indicates that cases of these two domains have relatively weak correlation regarding AI ethics.

3.3.3. Company scenario: HKGS-based analysis of cases within a specific company

In this section, Facebook, now renamed Meta platforms, and Amazon are chosen as examples for analysis by HKGS.

The method that averaged correlations are calculated is the same to that in Industry scenario. Fig. 20 shows the intrinsic correlation of Meta platforms cases.

As shown from Fig. 20, these cases have relatively strong correlation on affected groups, public attitude and response measures. These results are consistent with the facts that (1) affected groups in the AI ethics issues are mainly users on the company's social

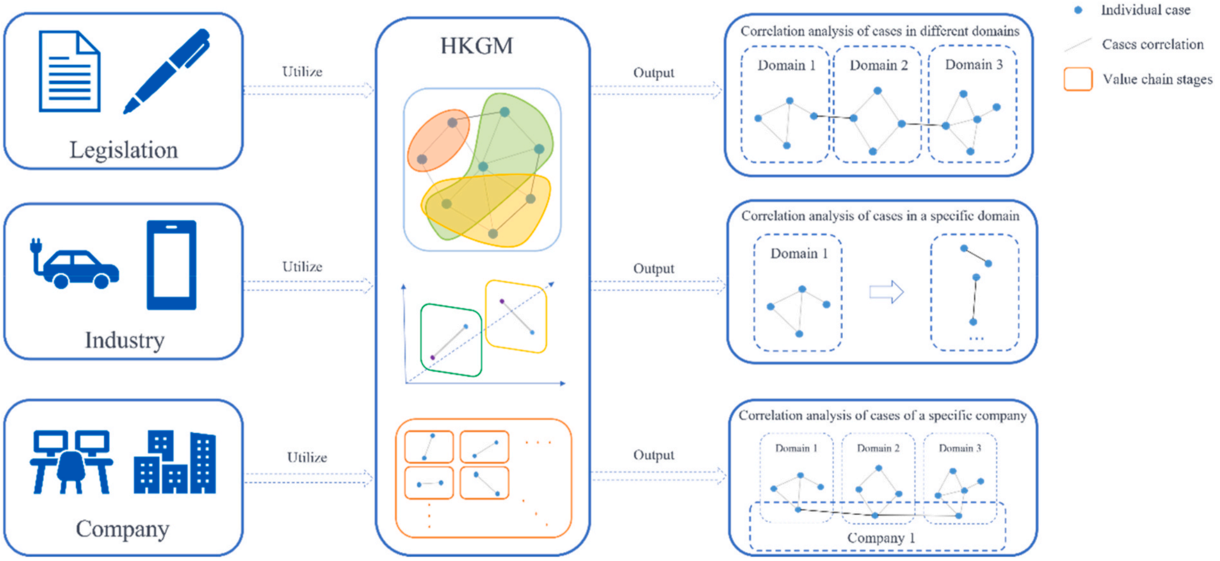


Fig. 14. Scenarios and their respective scopes.

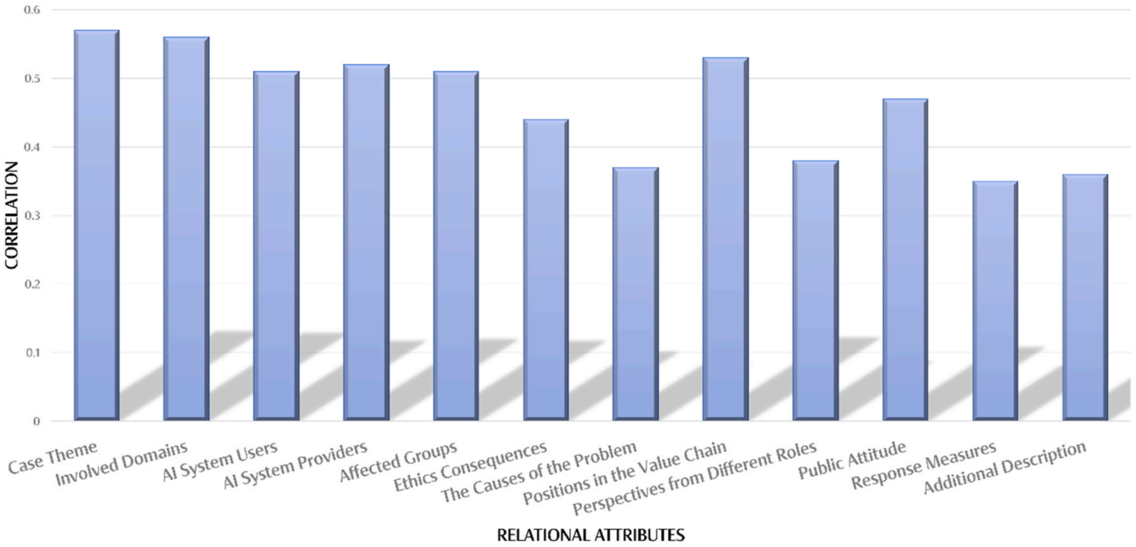


Fig. 15. The correlations between relational attributes of autonomous driving cases.

platforms like Facebook [35], (2) public attitude in general is very negative for this company's AI incidents, (3) when an AI ethics incident occurs, the company's response measure is likely to be an improvement of the AI algorithm. Meta platforms cases also have a strong correlation of positions in the value chain. Fig. 21 shows the HKGS output organized by value chain. Meta platforms cases represent relatively strong correlations in data acquisition, data access, data modeling and result learning stages.

In company scenario, practitioners can devise suitable governance measures for AI ethics issues based on the obtained results. For instance, Meta platforms can conduct user research focusing on highly concentrated affected groups. Undoubtedly, this significantly enhances the efficiency of addressing AI ethics issues. Then the company needs to be more proactive in addressing negative public attitude by providing more precise and convincing response measures. The company's data engineers should be vigilant in stages of data acquisition, data access, data modeling and learning result, and ensure that their behavior will avoid AI ethics problems.

3.4. Evaluation based on sample queries

In this chapter, the performance of HKGS is evaluated by setting up five sample queries at different levels corresponding to the query task of the knowledge system as shown in Table 5.

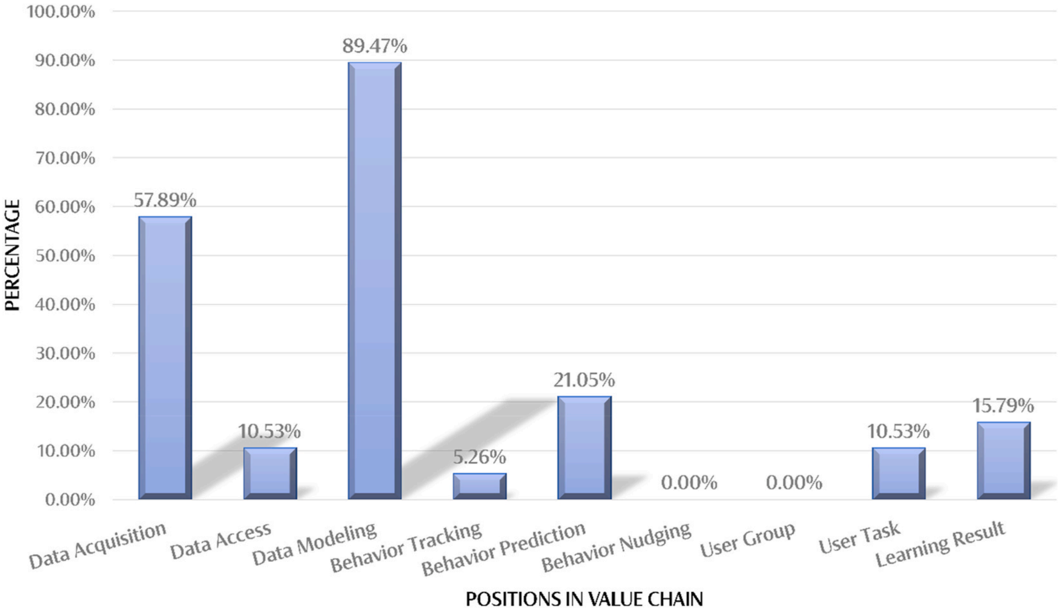


Fig. 16. The distribution of ethics cases on autonomous driving by positional attributes in value chain.

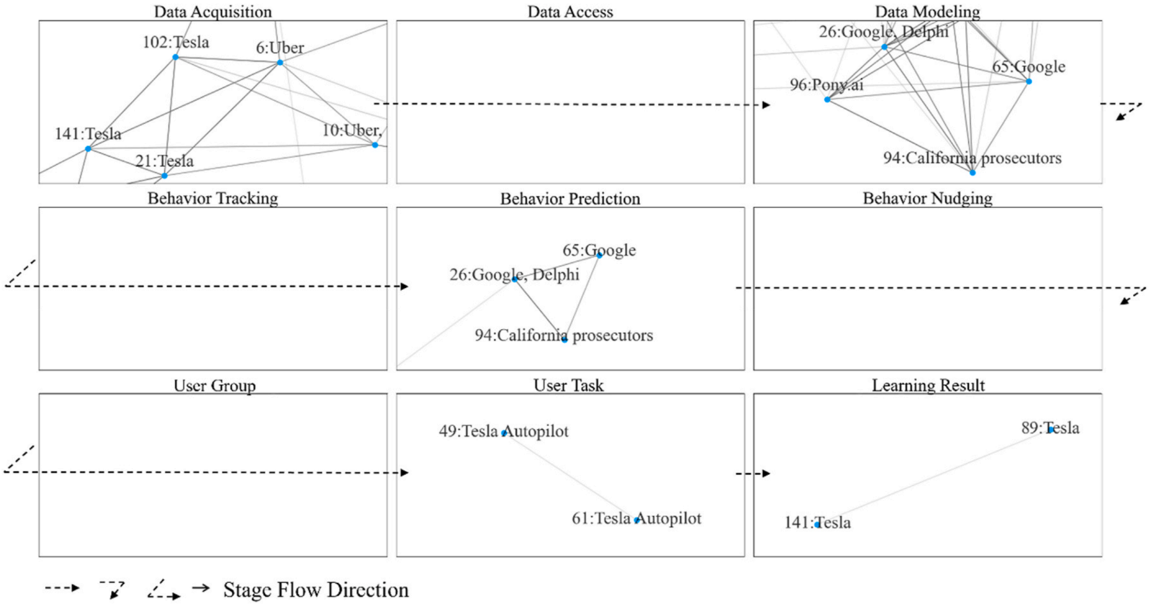


Fig. 17. The HKGS output of autonomous driving cases.

For the baseline, three other knowledge systems related to AI ethics are selected. First, a basic relational database is created. It is based on the Structured Query Language (SQL) for querying data. Second, a traditional graph database is constructed. It uses SQL-like statements to query the data. Then, a large language model (LLM) is created by fine-tuning the AI ethics cases database. Its data query is done through natural language. For HKGS, the data query is done through natural language-like statements. The final test results are shown in Table 6. Inside it, "✓" means that the system can handle the problem while "✗" means that it cannot.

From the results of Tables 6 and it can be concluded that HKGS has better data query capability. It can combine time, location and value chain information for correlation calculation and analysis. In addition, HKGS allows natural language-like utterance search, which is relatively straightforward to use.

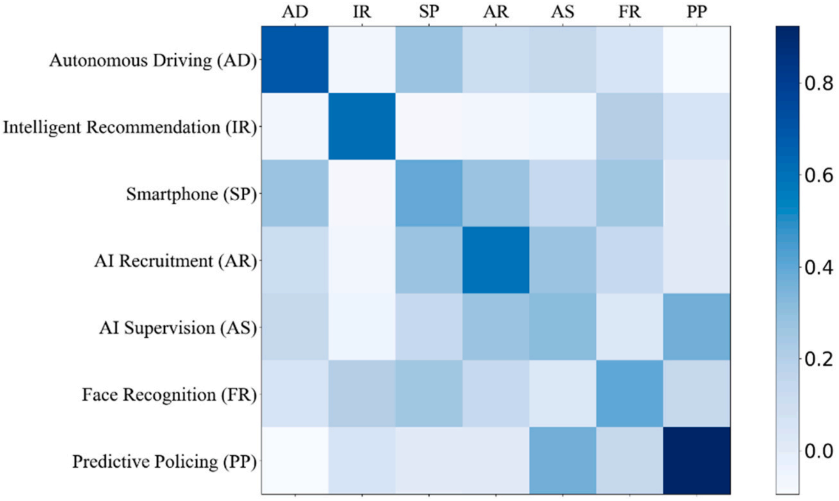


Fig. 18. The analysis of cross-domain correlations.

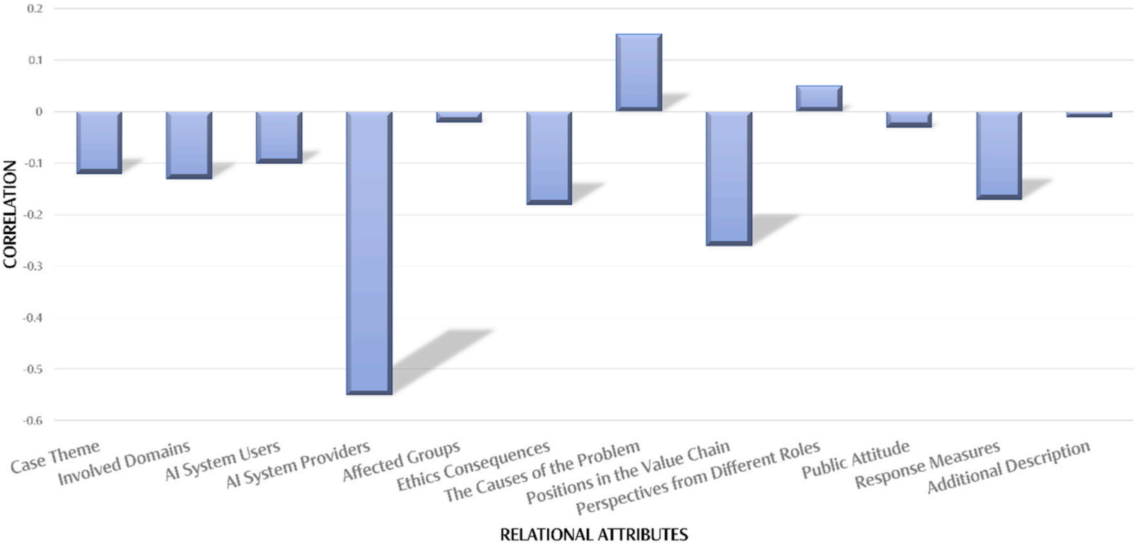


Fig. 19. The correlation between cases of autonomous driving and intelligent recommendations.

4. Discussion and conclusion

In this article, the authors proposed COOM and HKGS for the research of collected AI ethics cases. COOM provides the modelling criteria for the cases in the research from three perspectives, event attributes, relational attributes and positional attributes in the value chain. Based on COOM, HKGS provides a promising approach to combine the advantages of macro- and micro-level perspective and extend them.

Current approaches to the study of AI ethics are focused either on broad ethics guidelines or on specific domain. If the AI ethics is comprehended as a hierarchical tree structure from general concepts to special cases, these two research focuses represent a macro-level and a micro-level perspective respectively. That is to say, intermediate levels are missing. There is a lack of a coherent approach that connects all levels and enables a holistic understanding. In contrast to previous related studies, COOM pioneers the modeling of AI ethics cases from three perspectives. Due to the unique features of COOM, HKGS enables the analysis of cases within the same or across different domains. When combined with positional attributes in the value chain, these analyses can illustrate how correlation changes with the flow of stages in the value chain axis.

The results derived from the HKGS are predictive and interpretable. The predictive nature of the results can be attributed to the utilization of GCN by HKGS for link prediction within the knowledge graph. This facilitates the prediction of potential correlations among AI ethics cases. The interpretability of the results can be represented by the consistency between the results of the correlation analysis for specific attributes and the manual analysis in the aforementioned scenarios. Based on the demonstrated experimental

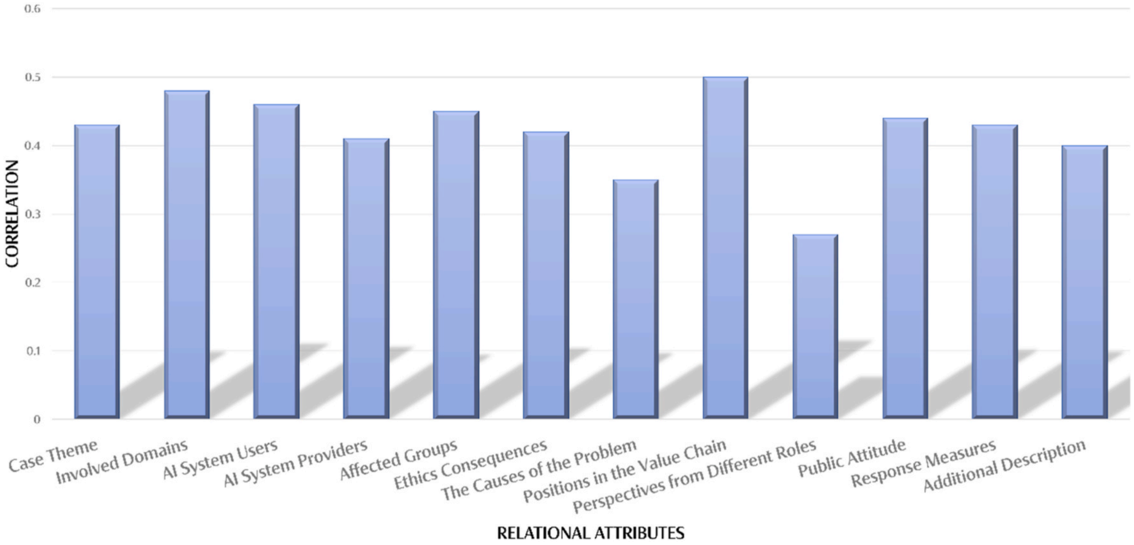


Fig. 20. The relational attributes' intrinsic correlations of Meta platforms cases.

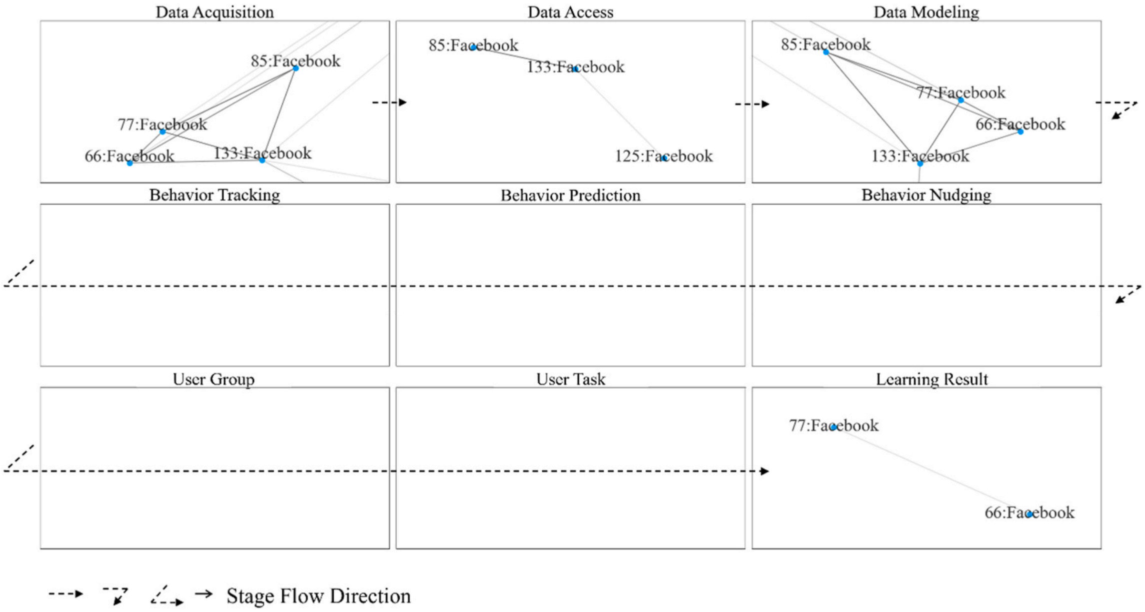


Fig. 21. The HKGS output of Meta platforms cases.

Table 5
The detailed description of sample queries.

Index	Query	Description
1	What are the AI providers for the case with index 5 ?	It is used to test the basic data query function.
2	Which location will have the most AI ethics cases in 2021?	It is used to test the ability to integrate location and time data
3	What are the cases whose relational attributes have the highest correlation with case 102 ?	It is used to test whether there is a single-case correlation calculation capability.
4	What are the cases whose relational attributes have the highest correlation with case 102 in data acquisition stage?	It is used to test if there is a capability to compute a correlation at a certain stage of the value chain.
5	Which relational attributes have high intrinsic correlation in the case of autonomous driving?	Used to test for the capability to compute intrinsic correlations for cases within a domain.

Table 6
The test results of sample queries.

Baseline	*Query 1*	*Query 2*	*Query 3*	*Query 4*	*Query 5*
Relational database	✓	✓	✗	✗	✗
Graph database	✓	✓	✗	✗	✗
LLM with fine-tuning	✓	✗	✓	✗	✗
HKGS	✓	✓	✓	✓	✓

results from HKGS, relevant practitioners in all three scenarios can obtain valuable insights tailored to their respective perspectives.

5. Limitation and future work

The research in this article is limited in three aspects. First, HKGS is a data-driven system. The actual dataset generation process is time-consuming. Second, there is also a requirement for further improvement in the visualization methodology of hypergraphs in HKGS. Third, the credibility of data sources and the feasibility of expanding data need to be substantiated.

To address these three limitations, there are three technical focuses for future work. First, the automatic labeling system based on LLM will be tried. In other words, the automatic annotation method needs to be attempted. Second, the visualization of the timeline of events can be improved to provide a better depiction of the evolution of hyper-knowledge graph. For example, dynamic display methods such as animation will be tried. Third, multi-source data collection methods must be used to avoid analytical bias from data homogenization. In addition, new parameters of credibility and feasibility of data extension need to be explored.

In addition, the authors have future plans to open source the entire system and publish it as a web page. It will also allow various practitioners in the AI domain to provide feedback or directly contribute new AI ethics data.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

Data will be made available on request.

Ethics declarations

All participants provided informed consent to participate in the study.

Review or approval by an ethics committee was not needed for this study because this study does not include any bioethically questionable data and is based entirely on open source datasets.

CRediT authorship contribution statement

Chuan Chen: Writing – review & editing, Visualization, Data curation, Conceptualization, Writing – original draft, Investigation, Methodology, Software, Supervision. **Yu Feng:** Writing – review & editing, Visualization, Supervision. **Mengyi Wei:** Visualization, Investigation, Data curation. **Zihan Liu:** Writing – review & editing, Visualization, Investigation. **Peng Luo:** Supervision, Investigation, Conceptualization. **Shengkai Wang:** Visualization, Methodology, Investigation. **Liqiu Meng:** Writing – review & editing, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Chuan Chen completed the theoretical formulation, data collection, experiments, experimental analysis, and article writing. Yu Feng was involved in data collection, experiments, and article revisions. Mengyi Wei supported in theoretical formulation, data collection and article revisions. Zihan Liu was involved in article figure drawings and article revisions. Peng Luo, Shengkai Wang and Liqiu Meng supported in theoretical formulation, experimental analysis. We would like to thank the researchers mentioned above.

References

[1] M. Wei, Z. Zhou, Ai ethics issues in real world: evidence from ai incident database, arXiv preprint arXiv:2206.07635 (2022).

[2] Y. Du, On the transparency of artificial intelligence system, Journal of Autonomous Intelligence 5 (1) (2022).

[3] S. Engelmann, M. Chen, F. Fischer, C.Y. Kao, J. Grossklags, Clear sanctions, vague rewards: how China's social credit system currently defines" good" and" bad" behavior, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, 2019, pp. 69–78.

[4] S.T. Segun, From machine ethics to computational ethics, AI Soc. 36 (1) (2021) 263–276.

[5] P. Fung, H. Etienne, Confucius, cyberpunk and Mr. Science: comparing AI ethics principles between China and the EU, 3, AI and Ethics, 2023, pp. 505–511.

[6] V. Dignum, Responsible artificial intelligence–from principles to practice, arXiv preprint arXiv:2205.10785 (2022).

[7] T. Hagendorff, The ethics of AI ethics: an evaluation of guidelines, Minds Mach. 30 (1) (2020) 99–120.

[8] A. Jobin, M. Ienca, E. Vayena, The global landscape of AI ethics guidelines, Nat. Mach. Intell. 1 (9) (2019) 389–399.

[9] M. Ryan, B.C. Stahl, Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications, J. Inf. Commun. Ethics Soc. 19 (1) (2020) 61–86.

[10] J. Morley, L. Floridi, L. Kinsey, A. Elhalal, From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices, Sci. Eng. Ethics 26 (4) (2020) 2141–2168.

[11] M. Coeckelbergh, Ethics of artificial intelligence: some ethical issues and regulatory challenges, Technology and Regulation 2019 (2019) 31–34.

[12] J. Whittlestone, R. Nyrup, A. Alexandrova, S. Cave, The role and limits of principles in AI ethics: towards a focus on tensions, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 195–200. January.

[13] M. Jantunen, E. Halme, V. Vakkuri, K.K. Kemell, R. Rousi, T. Mikkonen, P. Abrahamsson, Building a Maturity Model for Developing Ethically Aligned AI Systems, 2021.

[14] C.E. Prunkl, C. Ashurst, M. Anderljung, H. Webb, J. Leike, A. Dafoe, Institutionalizing ethics in AI through broader impact requirements, Nat. Mach. Intell. 3 (2) (2021) 104–110.

[15] D. Arnold, W. Dobbie, P. Hull, Measuring racial discrimination in algorithms, May, in: AEA Papers and Proceedings vol. 111, American Economic Association, Nashville, TN 37203, 2021, pp. 49–54, 2014 Broadway, Suite 305.

[16] U. Aïvodji, H. Arai, O. Fortineau, S. Gambs, S. Hara, A. Tapp, Fairwashing: the risk of rationalization, in: International Conference on Machine Learning, PMLR, 2019, pp. 161–170. May.

[17] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, W.Y. Wang, Mitigating gender bias in natural language processing: literature review, arXiv preprint arXiv:1906.08976 (2019).

[18] J. Zou, L. Schiebinger, Design AI so that it's fair, Nature 559 (7714) (2018) 324–326.

[19] B. Hutchinson, V. Prabhakaran, E. Denton, K. Webster, Y. Zhong, S. Denuyl, Unintended machine learning biases as social barriers for persons with disabilitiess, ACM SIGACCESS - Accessibility Comput. (125) (2020), 1-1.

[20] B. Baron, M. Musolesi, Interpretable machine learning for privacy-preserving pervasive systems, IEEE Pervasive Computing 19 (1) (2020) 73–82.

[21] B. Attard-Frost, A. De los Ríos, D.R. Walters, The Ethics of AI Business Practices: a Review of 47 AI Ethics Guidelines, AI and Ethics, 2022, pp. 1–18.

[22] B.C. Stahl, J. Antoniou, M. Ryan, K. Macnish, T. Jiya, Organisational responses to the ethical issues of artificial intelligence, AI Soc. 37 (1) (2022) 23–37.

[23] Jiazhi Xia, Jie Li, Siming Chen, Hongxing Qin, S. Liu, A review of research at the intersection of visualisation and artificial intelligence, Sci. China Inf. Sci. 51 (11) (2021) 1777–1801.

[24] N.K. Corrêa, C. Galvão, J.W. Santos, C. Del Pino, E.P. Pinto, C. Barbosa, E. Terem, Worldwide AI Ethics: a review of 200 guidelines and recommendations for AI governance, arXiv preprint arXiv:2206.11922 (2022).

[25] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, arXiv preprint arXiv:1609.02907 (2016).

[26] M. Haenlein, M.H. Huang, A. Kaplan, Guest editorial: business ethics in the era of artificial intelligence, J. Bus. Ethics 178 (4) (2022) 867–869.

[27] M. Jakesch, Z. Buçinca, S. Amershi, A. Olteanu, How different groups prioritize ethical values for responsible AI, in: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 310–323. June.

[28] Welcome, to the Artificial Intelligence Incident Database, 2023/07/29. https://incidentdatabase.ai/.

[29] A.F. Hayes, K. Krippendorff, Answering the call for a standard reliability measure for coding data, Commun. Methods Meas. 1 (1) (2007) 77–89.

[30] Incident 145: Tesla's Autopilot Misidentified the Moon as Yellow Stop Light https://incidentdatabase.ai/cite/145, 2023/08/01.

[31] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[32] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Adv. Neural Inf. Process. Syst. 26 (2013).

[33] Incident 20: a collection of Tesla autopilot-involved crashes. https://incidentdatabase.ai/cite/20, 2016/06/30.

[34] Incident 106: Korean chatbot luda made offensive remarks towards minority groups. https://incidentdatabase.ai/cite/106,.

[35] S. Frenkel, C. Kang, An Ugly Truth: inside Facebook's Battle for Domination, Hachette UK, 2021.