



**Trinity College Dublin**

Coláiste na Tríonóide, Baile Átha Cliath

The University of Dublin

School of Computer Science and Statistics

# External impact assessment on ethics of AI using ontology based on UNESCO recommendations

Udisha Chaudhary

August 19, 2022

A dissertation submitted in partial fulfilment  
of the requirements for the degree of  
MSc. Computer Science - Data Science

# Declaration

I hereby declare that this dissertation is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at <http://www.tcd.ie/calendar>.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at <http://tcd-ie.libguides.com/plagiarism/ready-steady-write>.

Signed:

Date:

# Abstract

Artificial intelligence has proved to be one of the most modern and revolutionary inventions in the history of humankind. With internet being widely available, everyone wants to advance towards automation. From giving less crowded directions on google maps to suggesting music on Spotify, AI indeed makes life much easier.

Recent studies have shown that AI has been the cause for many violations that includes privacy breach, biased results, excessive energy consumption, defying human rights and dignity. These intelligent algorithms are trained on historic data that might be skewed towards a particular gender, race or culture. Bodies like the EU have proposed their own principles on AI ethics. However researchers believe that such standards have to be designed in a way that it can be applied globally. UNESCO has come up with some recommendations for tracking AI ethics violations in the form of standards that a responsible AI should obey in all stages of its life.

Till now companies have been doing their own impact assessments which are rather partial towards their policies and agendas. However, it is equally important to monitor the actions of AI when it reaches the end users or is deployed in an environment.

This research aims at building an ontology named as AIIA-Ontology for the purpose of extracting and structuring the information in UNESCO document in order to use it as a model for conducting external impact assessment on a number of real world AI cases. This external assessment will help third parties that could also be end users to report an AI incident by examining its impact, underlying risks, and violations of set guidelines by UNESCO.

This report will explore the use of ontologies for representing any information in a domain of interest. Moreover it explains the entire development process of AIIA-Ontology along with few queries for information retrieval. It will also see a web application designed for conducting impact assessment.

By the end of this research it was learnt that ontology turned out to be an innovative and highly effective way for modelling UNESCO recommendations. The results of the impact assessment were also very satisfying in terms of AIIA-Ontology's ability to describe various AI incidents.

# Acknowledgements

I would like to thank my supervisor Professor David Lewis whose valuable inputs and guidance always got me into right direction during the course of this research.

I also want to thank my family especially my sister who kept supporting and motivating me throughout this research.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Research Questions . . . . .	2
1.3	Research Objectives . . . . .	2
<b>2</b>	<b>Literature Review</b>	<b>4</b>
2.1	AI Ethics - State of the Art . . . . .	4
2.2	An Initial study by UNESCO . . . . .	5
2.3	Ontologies and its applications . . . . .	7
2.4	Methods for building ontology . . . . .	9
2.5	Approaches to ontology evaluation . . . . .	10
<b>3</b>	<b>Design and Methodology</b>	<b>12</b>
3.1	Introduction . . . . .	12
3.2	UNESCO Recommendations on Ethics of AI . . . . .	13
3.3	Ontology development . . . . .	16
3.3.1	Scoping or Conceptualisation . . . . .	17
3.3.2	Reusing Existing Ontologies . . . . .	18
3.3.3	Determining Appropriate Core Terms . . . . .	18
3.3.4	Class and Class Hierarchy . . . . .	18
3.3.5	Properties . . . . .	23
3.3.6	Restrictions . . . . .	24
3.3.7	Individuals or Instances . . . . .	26
3.4	Ontology Encoding . . . . .	26
3.5	Querying Ontology . . . . .	26
3.5.1	Query 1 : Finding all the subclasses of ' <i>Risks</i> ' . . . . .	28
3.5.2	Query 2 : Finding all the subclasses of ' <i>Impact_zones</i> ' . . . . .	28
3.5.3	Query 3 : Finding all the instances of a given ' <i>Risks</i> ' . . . . .	29
3.5.4	Query 4 : Finding all the values or principles violated by a given instance of ' <i>Risk</i> ' . . . . .	30

3.5.5	Query 4 : Inserting data . . . . .	30
3.6	Web form for External Impact Assessment . . . . .	32
3.6.1	Impact assessment on an incident involving AI . . . . .	33
<b>4</b>	<b>Evaluation</b>	<b>36</b>
4.1	Context Based Evaluation . . . . .	36
4.2	Class Hierarchy Based Evaluation . . . . .	39
4.3	Quality Based Evaluation . . . . .	40
4.4	Application Based Evaluation . . . . .	43
4.4.1	First Iteration . . . . .	44
4.4.2	Second Iteration . . . . .	45
4.4.3	Third Iteration . . . . .	45
4.4.4	Database Evaluation . . . . .	46
<b>5</b>	<b>Conclusion</b>	<b>48</b>
5.1	Final discussion . . . . .	48
5.2	Future Work . . . . .	49
5.3	Final Reflection . . . . .	50

# Nomenclature

AI	Artificial Intelligence
AIAAIC	AI, algorithmic and automation incidents and controversies
OWL	Web Ontology Language
RDF	Ontology Language
AIIA	Artificial Intelligence Impact Assessment
LMICs	Low and middle income countries
LDCs	Least developed countries
W3C	World Wide Consortium
SDGs	Sustainable Development Goals

# 1 Introduction

## 1.1 Motivation

Today, Artificial intelligence has found its footing in every field of work and its full potential is yet to be explored as it is way more than what it seems right now. AI has already started accomplishing cognitive tasks it was built for and is inching closer to replacing humans at various places. For example, AI is capable of detecting a disease and even suggesting appropriate treatments. Radiologists in Austria have adopted an AI that detects tumours accurately whereas Denmark has come up with an AI system that facilitates faster treatment for urgent heart attack patients just by sensing the sound of the caller's voice (1). AI has gradually made a seamless presence in our lives starting from music recommendations on Spotify to self-driving cars. All of this attributes to powerful machine learning algorithms which are the backbone of any AI system. The catalyst that goes into the process of training these algorithms so they could make logical decisions like humans is the 'data'. The more the algorithms are fueled by data the better the results are. Another vital requirement of an AI is computer with massive energy and fast processing speed. Looking at the tremendous growth in AI, numerous companies and governments across the globe are coming forward to invest towards AI by providing computational strength and database needed for the execution of complex algorithms.

After everything, just like a coin, AI also has two sides. The data based on which the machine makes decisions often contains personal information about general public. This data is expected to be used and collected in such a way that it does not violate anyone's right to privacy and freedom. Moreover, people need to be aware of the purpose for which their data is being used. This is the reason why security and privacy alerts show up when downloading an application on mobile phones. Other concerns related to AI include enormous energy consumption and how much of impact they have on the environment. Many a time, AI systems have dual purposes which means they try to achieve more than the legitimate goals. Lastly and a very common problem is that the data used for training the algorithms is a reflection of today's situation. As a consequence, AI tends to



aggravate the biases that exists in society with respect to gender, race, color, ethnicity, etc.

On a good note, the world has started to pay attention to these issues by giving importance to the ethics of AI. Since past few years, many debates and suggestions have been emerged on how to monitor an AI starting from its development to deployment phase. The European Council of October 2017 stated that EU needs a sense of urgency to address upcoming trends such as AI without compromising protection over data, digital rights and guidelines for ethics (1). However the end users of AI are not limited to certain regions, they could be anywhere in the world where internet is accessible.

UNESCO is the first organisation to produce a general framework, consistent with international law, by recommending some regulatory standards that could help in addressing the ethical concerns of an AI globally. UNESCO also mentions the significance of conducting ethical impact assessment that identifies underlying risks as well as the impacts of the AI.

## 1.2 Research Questions

UNESCO has asked its member states to form an impact assessment model using the guidelines set for AI ethics. Although such assessments are carried by AI owners internally while developing the AI, but it is very limited and superficial as companies usually do not care about ethics while satisfying their set agendas for AI. Therefore this research will bring out an external impact assessment model to help third parties like civil right groups, researchers, social advocates or even end users to report an AI incident based on UNESCO's set standards. What this means is that only those ethical concerns will be covered that occur when the AI system goes into the market. The questions this research aims to answer are :

1. To what extent can ontology extract the information given in UNESCO document - "Recommendations on the Ethics of Artificial Intelligence" .
2. To what extent can the same ontology prove to be an effective model for conducting external impact assessment on the ethics of AI.

## 1.3 Research Objectives

In this research, an ontology will be designed that aims to formally represent UNESCO model for ethical AI. Then this ontology will be used as a model to perform external impact assessment on some real world violations done by an AI taken from AIAAIC repository. AIAAIC is a publicly accessible repository that contains summary of some

controversial AI incidents that made to news (2). Broadly, the impact assessment will assist a person in identifying areas that have been impacted, risks that an AI possess, and finally based on the risks the assessment will decide which UNESCO guidelines are violated. The main objectives of this research are outlined below

1. Analysing UNESCO's recommendations on AI governance.
2. Developing an ontology based on UNESCO guidelines step by step.
3. Retrieving information from the designed ontology.
4. Using the ontology as a model for conducting external impact assessment on real controversial AI provided by AIAAIC repository.
5. Updating and refining the ontology.

## 2 Literature Review

### 2.1 AI Ethics - State of the Art

With the growing concerns regarding ethics of AI, researchers have started studying the existing recommendations and policies published by various organisations all around the world. Gill and Germann (3) found that the private sector has not been proactive enough in reacting towards any violation of human rights caused by their products. Combined efforts of companies and governments needs to be put in detecting and mitigating any risk that can have a negative impact on humans. Further they have stressed that in governing AI a global approach must be taken in order to benefit every part of the world.

A. Jobin (4) analysed a total of 84 documents from multiple stakeholders who have shown their efforts in producing ethics guidelines on AI including government, non-government or private bodies like Google who itself earns a major portion from AI. A close analysis of those documents found that there was not even a single idea or principle pertaining to ethics that was common in the documents. However more than half of these documents appeared to be converging upon five major ethics principles - justice and fairness, non-maleficence, transparency, privacy and responsibility. Sustainability and solidarity are rarely mentioned in the ethics guidelines. This could be a problem as AI is known to possess high computational power and is heavily relying on environment for energy resources. Also their use has relentlessly caused job losses which means AI is failing to bring everyone together as a whole.

C. Roche (5) also studied the same 84 documents as A. Jobin but much more closely to see how inclusive these guidelines are specifically in terms of how Global South have been discussed and participated in the debate. Also to find out if women of the world are under-represented in those discussions. Firstly it was found that overall Southern nations have almost negligible publications which could turn out to be an issue if these citizens are not taken into consideration. The corpus was divided into three categories - NGOs, Public and Private. Then through a content analysis, different codes were used in order to find the existence of a particular topic, for example, 'sustainable', 'women', 'Africa'

and so on. The final findings reveal that sustainability is covered in almost 58% of public documents as compared to just 25% and 11% in NGOs and private sector documents respectively. Although NGOs and private sector cover Africa 17% and 13% percent respectively private sector on the other hand did not even touch this topic once. The paper states that such imbalance in debates and discussions could be 'concerning in the global discourse'.

## 2.2 An Initial study by UNESCO

A preliminary study was conducted by UNESCO (6) in order to put forth some ethical concerns related to AI like high scale job losses and physical or emotional harm caused to humans, with the belief that it will help build a universal framework thereby ensuring global well being and peace. After an inclusive analysis, the following ethical concerns fall exclusively within UNESCO's domain and are generally not included in the discussion.

1. Education - AI has led to 'labour displacement' as companies now have an eye especially for multi-skilled people who can take their AI game to the next level. Education department will be forced to go through a major shift from traditional ways to a tech-oriented curriculum that will train students to adapt to this advancing digital world. Moreover learning through internet reduces diversity of research and sometimes popular content overshadows the content that can be more relevant.
2. Culture and diversity - the advent of software like auto tune and deepfakes technology, has rendered creativity, personal skills and ingenuity rather unwanted. There are systems that can create and paint in 3D. AI also pose a challenge to diversity and inclusivity as not every artist out there possesses an easy access to AI technologies. Even the suggestions on Spotify and Netflix, which are majorly customised based on popularity and trend, restricts one from exploring other genres.
3. Science - An AI is optimised to produce solutions which are based on previous data of questions and search. This kind of problem solving algorithm deeply contrasts with the conventional ways used by the scientists and scholars incorporating detailed explanations, facts and formulas to solve the same problem. Introduction of AI to healthcare industry has automated the ways of diagnosis and treatments and is widely being used by people for self-diagnosis and cure. Moreover AI is also used for therapy nowadays. Although this is a huge revolution in health science, it can never replace the warmth of human care that only real health care professionals can provide. Lastly, they are also a source of massive amount of e-waste released in the environment.

4. Communication - Spreading misinformation or faux news is a common ethical impact of AI. UNESCO recalls the most controversial 'Cambridge Analytica' case to stress how artificial intelligence plays with human minds and has the power to develop and promote biased point of view and mentality in the society. AI has also proved its benefits in the field of journalism. It can produce articles on its own based on the available information but the job cannot be entrusted entirely to the AI as spreading content without human check can lead to false information and even cause tension.
5. Gender Equality - AI can also be customised in certain ways that may amplify the existing gender gap in the society. There was a report that Amazon had an AI trained to short list only male applications for the job hiring. The reason behind this was that the AI was using algorithms trained on previous candidates data that had male candidates in majority.
6. Peace and security - If AI is allowed to make military decisions, it could lead to war tensions and exploitation of enemy. This would violate UNESCO's agenda of promoting world peace and harmony.
7. Challenges for African Nations -Southern countries of the globe that are still developing generally run short on infrastructure, electricity generation and business development. They still have limited access to AI technologies while the rest of the world is using AI on a daily basis. The growth in AI technology is also not balanced as China and North America together account for 70 percent of AI economic impact in the world.

These are only a few points from the entire study by UNESCO on AI ethics. It sums up why UNESCO is interested in AI ethics and who will be benefited by their work. Moreover, the above issues hinder with the goals of the United Nations 'Sustainable Development agenda' (8) for the year 2030 that talks about equality among countries, gender, cultures, sustainable consumption of resources and worldwide economic growth.

After inspecting the situation and consulting multiple stakeholders who themselves could be impacted by AI in some ways, UNESCO adopted the world's first globally applied recommendations and policies on the Ethics of AI on November 24, 2021. Those recommendations, analysed in the subsequent chapter, exclusively address the above listed areas of impact.

## 2.3 Ontologies and its applications

This section introduces ontologies and its various applications with an intention to justify their use in this research paper. An overview of important terms that make up an ontology is also provided along with some examples where ontologies have been implemented to represent some sort of knowledge.

The first paper that was found to have a detailed introduction to ontology belongs to T. R. Gruber (9). It states that for any information that needs to be formally represented, having a common shared vocabulary that represents concepts, entities and their relationships can be quite helpful. This is basically called conceptualisation of a given area of interest. Ontology is an explicit form of that conceptualisation. Another appropriate definition given by Arbanas and Cubrilo on ontology that fits perfectly well with the scope of this dissertation is (10)

"An ontology is a specification of concepts and their relationships which represents knowledge in a formal and structured format and provides a better tool for communication, reusability and organisation of knowledge."

M. Uschold and M. Gruninger (11) take a different approach for defining ontologies in their paper. They enumerate some major uses of ontologies. Some of them are :

1. Communication : ontologies help in communicating any information in an effective way leaving no ambiguities. They provide a unifying model that cover all the necessary definitions of concepts and their relations with each other.
2. Interoperability : sometimes two people from different languages using different software platforms may use a common ontology to exchange data. In such cases ontologies help in translating the shared concepts into more than one language.
3. Reusability : ontologies are designed so that they can be shared and integrated in numerous software applications. A class of one domain ontology can be used in another.

All of the above mentioned uses can help in spreading UNESCO's framework to larger audience on internet and can fulfill the aim of building a universal AI monitoring model. This next paper written by Noy and Deborah (12) is highly recommended by ontology engineers. It explains all the fundamental elements required to build an ontology such as 'classes' also called concepts in the domain of interest. Then there are few 'properties' to define various features or relations of a concept. Moreover a class can be divided into 'subclasses' for more specific conceptualisation. These elements as well as few others will be more elaborated while developing the ontology in the subsequent chapters.

World wide web consortium(W3C) has another term for ontology, called 'vocabulary' and defines it as a collection of concepts and relationships that describe and represent an area of study (13). Vocabulary goes from simple to complex as the number of relationships increases. These complex vocabularies are what they call ontologies. As N. Kamel (14) states, ontology can be a simple catalog of just terms in the knowledge domain or it can be complex when properties are included.

As for the applications, the biggest ongoing project on semantic web by W3C uses ontology (15). It is an initiative to basically link all the data on the web in order to make searching more accurate. K. Palaniammal (16) describes the role of ontologies in semantic web. Ontologies formally represent information on world wide web (www) using concepts and relationships. Machines use ontologies to find a specific information on the Internet. Benefits of using semantic web includes speedy response of search engines as well as getting the desired results that are more relevant to the query.

A research paper by Tania T. (18) on ontology engineering found that the applications of ontology covers a wide range of areas other than semantic web, for example, finance and medicine. Financial Industry Business Ontology (FIBO) (19) a model crafted to establish a common comprehensive understanding of ambiguous finance terms and concepts. Robert Stevens (20) explains how ontologies can be utilised by biologists to define biological data and the underlying concepts more concretely and in organised fashion. For example describing molecules' nature and interactions with other entities like cells.

On the whole, ontology can be interpreted as a knowledge supply which provides a better understanding of concepts. It is also clear that applications of an ontologies are not limited to only semantic web. It can be used to formally represent any domain of interest. Therefore for this research ontology seems to be a suitable tool for representing information on AI ethics.

## 2.4 Methods for building ontology

N. Kamel says, it is impossible for an ontology to model every piece of information in the domain. It should always target a certain amount of knowledge and the concepts within (14). Also, there are no clear conventional methods for implementing ontologies. But there are few practitioners whose work has been suggested widely and deemed as a good approach to building ontology. One of those works belongs to Noy and Deborah (12) as their paper aims at developing an ontology from the beginning that conceptualises wine and food knowledge. They specified a number of steps that can yield a good quality ontology. The most valuable suggestion they gave is that one must always start by scoping the ontology which means the developer should first strictly decide what the ontology is trying to achieve and select only the relevant concepts from the entire domain. Then start defining the classes and their properties. These steps will be elaborated in the next chapter. M. Uschold and M. Gruninger (11) Also draws a very similar approach as Noy. Their recommended steps are :

1. Defining the scope and purpose of ontology
2. Creating an ontology which includes three sub steps i) ontology capturing that includes identifying fundamental concepts and their relationships with each other. ii) Ontology coding, which means explicitly representing concepts and relations using coding languages. iii) Integrating other ontologies if required.
3. Evaluation of ontology - to ensure the ontology is accurate and correctly implemented.
4. Documentation - this should provide users with a document explaining all the main concepts defined in the ontology and their meaning.

Ontologies can be interpreted by humans as well as software applications. According to W3C, a formally coded ontology enable machines to understand the conceptualisation of the domain of interest. OWL language is the programming language of an ontology that is widely accepted by engineers as well as semantic web. It contains predefined syntax for each element of ontology such as classes, subclasses, properties etc (17).

The concept of ontology has seen some innovative developments in applications that allow building ontology faster and without manually coding. One such application is Protege, created by Stanford University, that provides tools for adding components like classes, properties, subclasses (21). Another quite useful tool is Chowlk converter which allows one to conceptualise ontologies in the form of UML diagram by using a given set of templates. It then converts those elements in diagrams into OWL language.



## 2.5 Approaches to ontology evaluation

Evaluation is a fundamental step in finalising ontology before it gets shared and reused by others. The whole point of knowledge sharing fails if an ontology does not provide the desired results. Gómez-Pérez says currently ontologies are not that large and their practical uses are still not explored fully. As a result, very little progress have been made in building evaluation methods (22). Gómez have outlined certain evaluation criterias that require one to do a number of checks. Here are some of those checks that can be considered for this research

- Checking the structure of ontology - this is to ensue that the architecture is well built and obeys the rules of other ontology development platforms.
- Checking the syntax - this will assist in finding out any errors related to syntax such as incorrect keywords or no loops present.
- Checking the content - this evaluation makes sure anything the ontology defines is correct and consistent. It has to pass the test for three C's - Consistency, Completeness and Conciseness.

Consistency means there has to be no contradictory definitions either specified or inferred. Completeness refers to the level of coverage over the information while also taking account the degree of granularity set for ontology. If nothing is left to define then it can be called as complete. Finally, conciseness can be achieved by not adding any redundant information either explicit or derived. Conciseness is also based on how precise the definitions are, for instance if the properties of each class has been defined.

A survey on evaluation methods for ontology was done by J. Raad and C. Cruz (23) because they think it is imperative to validate them in order to use as reference models. Based on their analysis, in addition to the 3Cs introduced by Gómez-Pérez (22) an ontology must also fulfill these additional criteria

- Accuracy : this is to check the correctness of definitions of classes, properties and instances.
- Adaptability : it measure the extent to which an ontology can be put to use.
- Clarity : as the name suggest, how clear the ontology is in terms of communicating information. Definitions should be as general as possible in order to reach out bigger audience.
- Computational efficiency : this is to ensure that users get the desired speed in the task that the ontology has been put on.

Apart from criteria based evaluation, there are few other approaches that are suggested in

the said survey. For example comparing one ontology to another gold standard ontology that covers the same topic of interest. But this kind of approach is very challenging as it is very hard to find an ideal gold ontology. An alternative to this would be getting a group of human experts to make a set of important concepts or vocabulary that can then be used as gold standard. This modified way will be considered as one of the evaluation approach in this paper. Second recommended evaluation which is worth knowing is measuring the performance of an ontology when used for a specific task. This kind of application based approach is adopted in this paper as well.

A paper by a group of ontology engineers (24) identifies more than twenty bad practices in the ontology modelling which according to them is hard to detect. This paper is very informative in terms of understanding and mitigating errors that appear while development. These errors called pitfalls in the document are also classified based on the three Cs criteria. So it helps in knowing how each pitfall affects which criteria.

# 3 Design and Methodology

## 3.1 Introduction

This chapter explains in detail each step that has been taken to extract the information from the UNESCO document and formally represent it in the form of ontology. Future, it will also explain how the ontology is used as a model for carrying out external impact assessment. From here on, the ontology will be referred as AIIA-Ontology (Artificial Intelligence Impact Assessment Ontology). It is designed in such a way that it helps in annotating real world cases involving AI (from AIAAIC repository) in terms of impact and violations. The list given below gives an overview of what each section will explain. The sections are sequenced in the order in which the research took place for better perusal of the reader.

- **UNESCO Recommendations on Ethics of AI**

This section provides an analysis of UNESCO's document on the ethics of AI.

- **Ontology Development**

This section explains how an initial draft AIIA-Ontology is designed and developed. It will include all the necessary knowledge required to create a novel ontology.

- **Ontology Encoding**

In this step the ontology is formally encoded in OWL programming language.

- **Querying Ontology**

Here, few important SPARQL queries will be shown along with the results when executed over AIIA-Ontology. These queries help in retrieving information for external impact assessment on the ethics of AI.

- **Web Application for External Impact Assessment**

Here, a detailed description of how an impact assessment is conducted. A web application will be introduced which is designed to provide a platform for the execution and generation of the final results of assessment.

## 3.2 UNESCO Recommendations on Ethics of AI

One of the main purposes of AIIA-Ontology is representing UNESCO's framework on AI governance to have a common understanding among people of any background. Section 2.1 talked about the initial study done by UNESCO on the ethics of AI from their point of view and how it impacts their domain - education, culture, communication and science. Based on the findings, UNESCO released a document called "Recommendations on the ethics of Artificial Intelligence" (7) which contains guidelines for promoting responsible AI. It first starts by discussing the potential areas that could have a negative impact by an AI such as human minds for example decision-making, environment, human dignity, freedom and rights, gender equality and society. The ones that have been exclusively mentioned are the countries like LMICs (low and middle income countries), LDCs (least developed countries), etc as they are generally not given much attention. Some zones that have been recently identified as being hit by AI are healthcare, media, personal data or even democracy. Since applications of AI are now in every sector of work, their ethical implications are increasing concerns. As a result UNESCO has recommended the concerned authorities to bring all the stakeholders from both public and private sectors for example civil society groups, researchers, media, human rights and equality bodies, people from different age groups, companies and policy makers so that each opinion contributes in forming standards to govern AI and mitigating their risks.

The document chalked down all the relevant values and principles in order to ensure a standard behaviour from a responsible AI. UNESCO describes values as the ideal behaviour standards while the principles are more specific and will be used to operationalise the set of values in policies and actions. Under each value or principle, UNESCO addresses various kinds of risks or concerns related to an AI that could lead to some negative impact. UNESCO has suggested AI actors, whether they are users or legal authorities, to use these guidelines as a base for ethical impact assessment that should continue for the entire life of AI. While remembering the goal is to develop an external assessment model which is conducted when AI had already reached to the end users. Also at this point it is worth mentioning only those risks that pose serious threats to AI users or the environment where it is deployed are taken into account. Table 3.1 lists each value set by UNESCO along with the risks that can lead to its violation.

Table 3.1: UNESCO's Values and some risks that violate them

Values	Risks
Respect protection and promotion of human rights and fundamental freedoms and human dignity	<ul style="list-style-type: none"> <li>• Harming or subordinating human based on race, colour, gender, age, language, religion, opinion, nationality, ethnic background, disability, conditions of birth and or economic</li> <li>• Abusing human rights and freedom</li> <li>• Objectifying humans</li> </ul>
Environment and ecosystem flourishing	<ul style="list-style-type: none"> <li>• Leaving carbon footprints</li> <li>• Unsustainable exploitation and use</li> <li>• Deteriorating environment and ecosystem</li> </ul>
Ensuring diversity and inclusiveness	-
Living in peaceful, just and interconnected societies	<ul style="list-style-type: none"> <li>• Segregate, objectify and undermine freedom</li> <li>• Affects autonomous decision-making</li> <li>• Turning individuals and communities against each other</li> <li>• Threatening coexistence between humans, other living beings and environment</li> </ul>

Similarly table 3.2 represents the outlined principles and the risks owned by an AI that can impact them. It is important to realise that there are many other risks that an AI can have but the aim of this paper is to focus only on what UNESCO has addressed. Moreover, the external impact assessment developed in this research identifies momentous concerns only and due to this reason some value and principles have no risks mentioned against the them. These are although desirable but does not address any serious risks or threats. Lastly, all of the risks mentioned in the tables were identified manually after reading each value and principle in the proposed document. Since manual inspections are subjective, there is a chance of missing out on one or two. However, the essence of ontologies is that they are extensible which implies any user will be able to reuse and extend the AIIA-Ontology in their own way.

Table 3.2: UNESCO's principles and some risks that violate them

Principles	Risks
Proportionality and do no harm	<ul style="list-style-type: none"> <li>• Exceeding legitimate aims or objectives</li> <li>• Harming human beings, human rights and freedom, communities and societies</li> <li>• Takes critical irreversible decisions without human involvement</li> <li>• Life and death decisions</li> <li>• Used for mass surveillance</li> </ul>
Safety and security	<ul style="list-style-type: none"> <li>• Unwanted harms</li> <li>• Vulnerable to attacks</li> </ul>
Fairness and non-discrimination	<ul style="list-style-type: none"> <li>• Unavailable and inaccessible for certain people or countries - age groups, cultural systems, people with disability, women, different language, disadvantaged, marginalised and vulnerable people</li> <li>• Benefits not equally shared</li> <li>• Discriminatory or biased outcomes</li> <li>• Exacerbates digital and knowledge divides</li> </ul>
Sustainability	<ul style="list-style-type: none"> <li>• Hinder SDGs of UN</li> </ul>
Right to Privacy and Data Protection	<ul style="list-style-type: none"> <li>• Privacy invaded</li> <li>• Unethical data collection, use, sharing, deletion</li> <li>• Processing personal data without consent</li> </ul>
Human oversight and determination	<ul style="list-style-type: none"> <li>• Takes life and death decisions</li> </ul>
Transparency and explainability	<ul style="list-style-type: none"> <li>• People not informed about AI</li> <li>• No access to datasets or code</li> <li>• Unexplainable algorithms</li> </ul>
Responsibility and Accountability	-
Awareness and Literacy	-
Multi-stakeholder and adaptive governance and collaboration	-

### 3.3 Ontology development

It was learned in section 2.3, one of the applications of ontologies is communication (11) because they can be a medium through which knowledge can be represented formally to help establish a common understanding of concepts among people. Having an ontology dedicated towards UNESCO's adopted recommendations will help in generalising the concepts of AI ethic and form a common vocabulary to help communicate that across all stakeholders. Once this ontology is well conceptualised, it will serve as a model for classifying AI incidents as part of the external impact assessment. This section will describe each step used in building AIIA-Ontology.

A guide written by Noy and McGuinness (12) is followed throughout the process of building AIIA-Ontology. It is one of the most recommended papers that contains detailed instructions for beginners to develop ontologies from scratch. Another quite insightful research dedicated to designing and implementing ontology is by Uschold and Grüninger who also attempt to approach in similar ways (11). Unequivocally both these papers make a good reference point.

#### Key Elements

The following listed elements are the basic building blocks of an ontology. Hence understanding them is crucial as they are incorporated in AIIA-Ontology. Their description has been drawn from the guide itself (12).

1. Class or Concept : any entity that represents a concept in a domain of interest is classified as class.
2. Properties or Slots: these are the attributes used to describe classes. Properties are also used to define relations between two classes.
3. Instances : these are the individuals that are members of a class and inherits all of the properties of that class. Instances are introduced when the ontology reaches at the desired granularity.
4. Restrictions : these are basically some constraints or conditions that provide more information about classes and properties.
5. Namespace : the second main application of ontology is resusability. This is achieved by assigning them a unique resource identifier (URI) that supports reuse, extension or integration of one or more ontologies. In case a user decides to amend an existing ontology then a different URI has to be assigned to the latest version. Therefore each version of the ontology is stored uniquely.
6. Subclass : any class or concept can be further broken down to smaller concepts

or sub-classes. A subclass inherits all the properties of its parent class.

7. Knowledge base - all of the classes, properties and instances together form a knowledge base.

Based on the literature, there are no stringent rules to build ontology. It is usually up to the designer to figure out ways that work the best. Traditionally ontologies are finalised after going through a number of iterations (12). Henceforth, an initial draft of AIIA-Ontology covering all the fundamental concepts is developed in this chapter. Later in the evaluation part, this draft will go through a series of revision and refinement to achieve the final product.

### 3.3.1 Scoping or Conceptualisation

The first step in the guide (12) is scoping of the ontology that involves setting a goal and fundamental concepts required to achieve it. Here the designer should reflect on the purpose that needs to be fulfilled by the ontology.

The goal of AIIA-Ontology is to extract all the necessary information provided by UNESCO document about AI ethics in such a way that it can be used to perform external impact assessment on some controversial AI system.

The rationale behind the AIIA-Ontology is explained with the help of a diagram in fig 3.1. The blocks represent core concepts and the arrows are the properties that relate each concept to one another. There is an AI system, an organisation that owns it, any area that has been impacted negatively, concerns or the underlying risks owned by AI and finally the UNESCO's values and principles being violated by those risks. This design fulfills both the purpose of extracting the concepts from UNESCO's recommendations as well as capturing an the AI system impact and underlying issues.

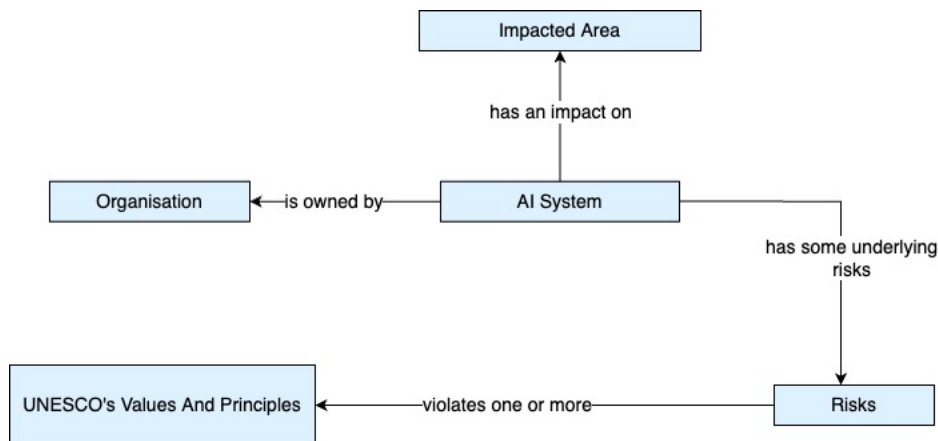


Figure 3.1: Rationale behind AIIA-Ontology



### 3.3.2 Reusing Existing Ontologies

As per the latest knowledge there is no ontology based on UNESCO's recommendations for AI ethics. The ontology that is closest to the AIIA-Ontology is called AIRO by Delaram Gopalya (25). It is designed for the purpose of 'internal risk assessment' which equips AI owners with a model to identify their product as a high-risk, based on AI Act and ISO 31000. It can definitely provide a good starting point as the core concepts are very similar, for instance AIRO contains Risks, Impacts, Events, AISystem. Amongst other models that are a bit similar includes V. Agrawal's ontology conceptualising ISO guidelines for risk management in the information security domain (26). Another work on information security is presented by Almut H. (27). It also assist in understanding the use of key terms in an ontology such as class, property and particularly restrictions.

### 3.3.3 Determining Appropriate Core Terms

Considering the scope of AIIA-ontology and using AIRO (25) as reference, the core terms that must be present are 'AI system', 'Owner' or 'Organisation' , 'Impact' of the AI, underlying 'Risks', UNESCO's 'Values' and 'Principles' to govern the AI ethics.

### 3.3.4 Class and Class Hierarchy

After setting a scope and identifying the essential terms, this is where the actual structure of the ontology starts. To begin, each concept takes the form of a class and if there is a scope of refinement, subclasses will be introduced. Before describing each class, it is important to set a naming convention so as to avoid any confusion between different elements of ontology. As per the guide (12), each concept name starts with a capital letter and if the name contains more than one word then each word should be separated by an underscore or space. Generally not every online platform supports space, hence underscore is used here.

#### *Chowlk Converter*

Before defining classes, it is important to introduce a tool called Chowlk. It is always considered a good strategy to first draw UML diagrams in order to achieve the desired piece of work. Chowlk gives users the freedom to conceptualise ontologies just like any other UML diagram consisting of entities and relationships (28). It comes with predefined templates for visual representation of elements in an ontology, such as class, subclass, properties, etc Chowlk library is first loaded into Google's diagrams.net application and once the diagram is finished, an XML file can be downloaded. Chowlk converter (29) identifies the elements by the help of templates and converts the entire diagram into OWL which is one of the programming language for coding ontology. The newly generated

OWL file can then be opened using any ontology developing platform like Protege to add more details and complexities.

As mentioned, each ontology is assigned a namespace URI that uniquely identifies it. Therefore a namespace for the ontology is defined using Chowlk template (see fig 3.2 ). Each individual element in an ontology is declared by appending its name to the URI.

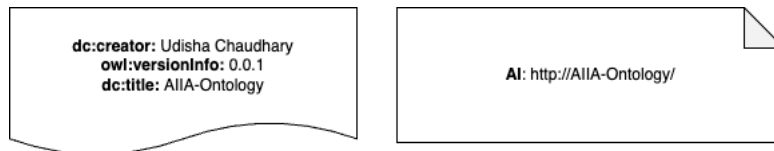


Figure 3.2: Defining URI for AIIA-Ontology using Chowlk templates

Now, here are the classes and subclasses defined in the AIIA-Ontology.

### **Ai\_System**

This class describes the AI system for which the impact assessment is being done. It will contain information like the name of the AI, the incident associated with it, a link that redirects to its corresponding page on the AIAAIC repository.

### **Organisation**

It is imperative to know whom the AI system belongs to. The reason this entity is given the designation of a concept or class is that it can be broken down further into private organisation and government organisation, each having their own subclasses and properties. However this paper will restrict till 'Organisation' concept only.

### **Impact\_zones**

Section 3.1 highlighted some of the areas or zones an AI system has some negative impact on according to UNESCO study. Therefore this class identifies various impact zones. It will help in bringing together stakeholders from these zones that are significantly affected. An ontology cannot include everything but a subset of information (12). As a result twelve zones have been selected from the UNESCO document while keeping in mind that there could be more. The list below provides a brief explanation for all the zones enumerating what aspects they cover that makes them a valid subclass.

1. `Employment_or_labour` : this subclass can be used to address different kinds of sectors where the employment is usually affected.
2. `Human_privacy_safety` : this is created to include all the governing bodies that work for maintaining privacy and safety of humans. It can also include various international laws that are based on ensuring human privacy and safety as these legal frameworks are also impacted by the AI incident.

3. Human\_rights\_freedom : it includes governing bodies that ensure people's rights and freedom as well as different kinds of international law that talk about human rights and freedom.
4. Culture : this class can further classify different cultures and diversity groups.
5. Environment : this class encapsulates categories like biodiversity, other living organisms, ecosystem. It also takes into account the goals of SDGs set by UN.
6. Health\_care : this concept is deemed worthy to be added as according to UNESCO it has emerged as a latest zone of impact.
7. Gender\_equality : it can be used to involve groups and organisations that are constantly working for women empowerment.
8. Education\_or\_future\_generation : it mainly focuses on the sustainability aspect but can be used to address education as well.
9. Marginalisation\_or\_vulnerable\_people : this category is very important from UNESCO point of view. It can be used to identify old people, children, people with some disability and minorities that gets neglected by technology.
10. Communities\_and\_groups : another very important concept that helps in identifying people of different race, caste, color, nationality, opinion, language, religion, different ethnic background.
11. Media : it covers social media, journalism, or online and offline platforms.
12. Countries : this is to address LMICs, LDCs, and countries which are not so advanced in technology or digital world.

The diagram shown in fig. 3.3 gives an illustration of how a class and its subclasses are created using Chowlk library's templates. So the choice of boxes and arrows are in sync with the notations described by Chowlk.

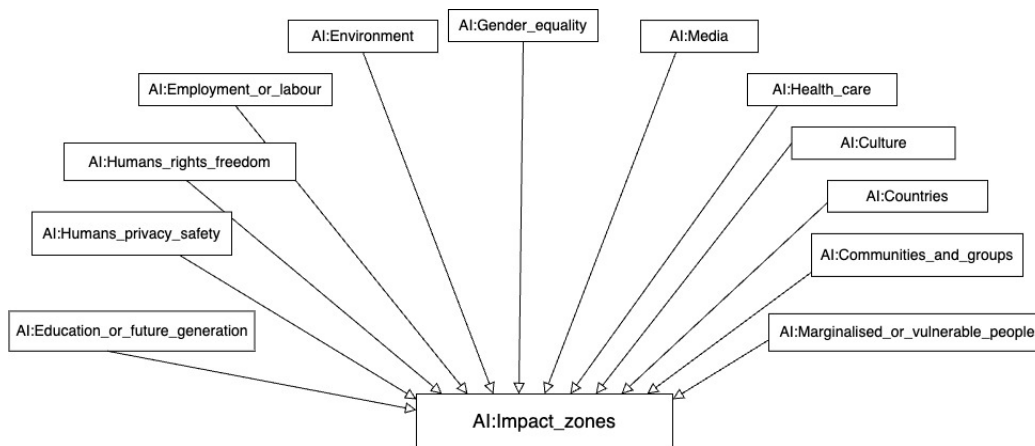


Figure 3.3: Defining '*Impact\_zones*' and its subclasses using Chowlk template

## Values

It is one of the most important class that equips the ontology with all the values set by UNESCO. Each sub-class of this concept represents a value mentioned in the table 3.1. While UNESCO considers each value desirable, however, out of all the following three values are incorporated in AIIA-Ontology as they were found to be addressing distinct issues.

1. Respect, protection and promotion of human rights and fundamental freedoms and human dignity
2. Environment and ecosystem flourishing
3. Living in peaceful, just and interconnected societies

The diagram in fig. 3.4 illustrates how '*Values*' class along with the subclasses is created using Chowlk.

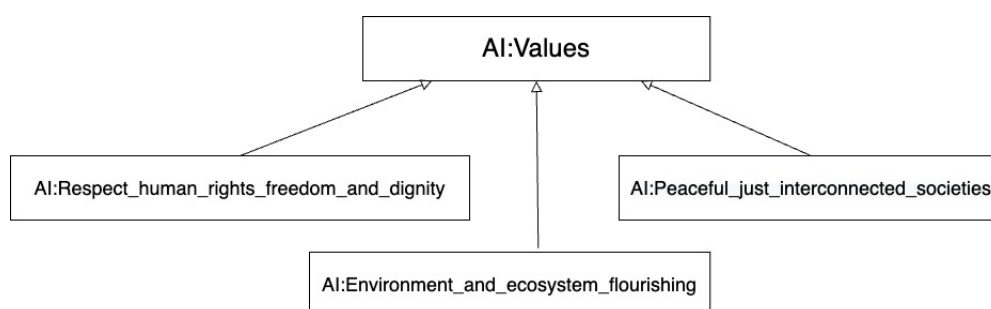


Figure 3.4: Defining '*Values*' and its subclasses using Chowlk template

## Principles

This class includes all the necessary principles picked up from the table 3.2 Here are the selected principles deemed worth of adding into AIIA-Ontology.

1. Proportionality and Do No Harm
2. Fairness and non-discrimination
3. Safety and security
4. Right to Privacy, and Data Protection
5. Human oversight and determination
6. Transparency and explainability

The diagram shown in fig. 3.5 shows 'Principles' class together with subclasses.

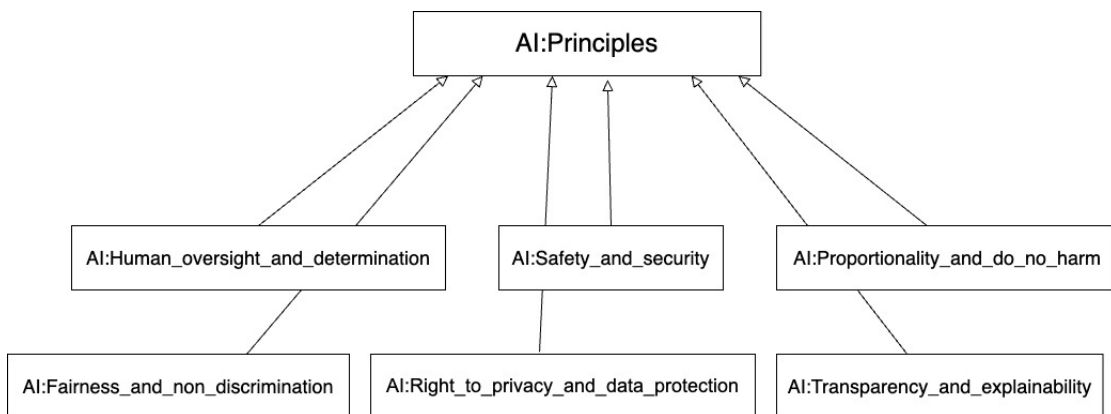


Figure 3.5: Defining 'Principles' class along with its subclasses

## Risks

This class represents the risks of an AI system. It is divided into specific categories covering all the concerns listed in table 3.1 and table 3.2.

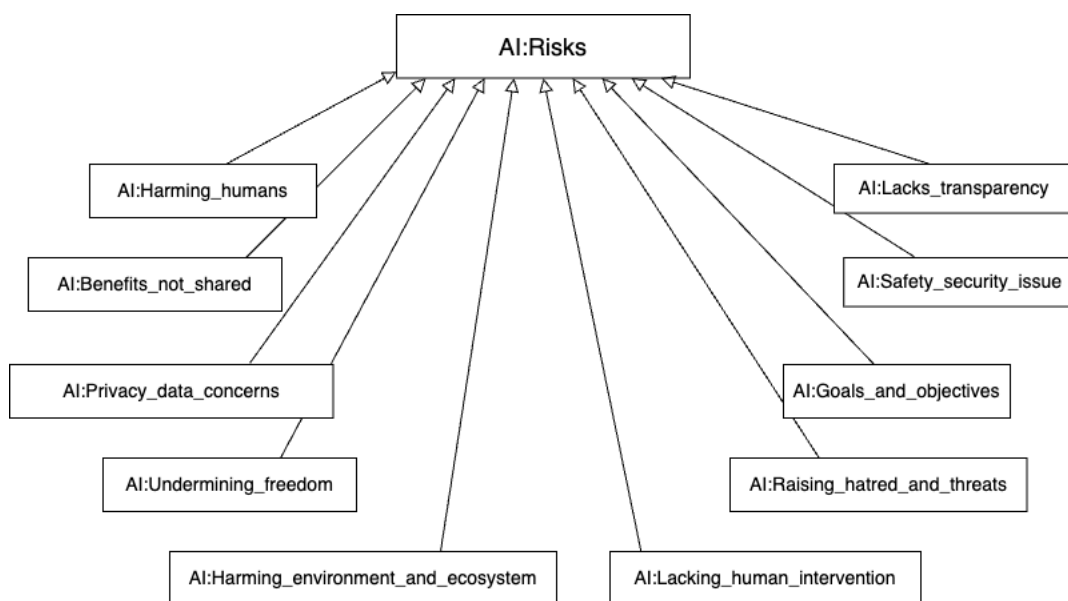


Figure 3.6: Sub classes of 'Risks' using Chowlk template

### 3.3.5 Properties

Defining classes is an important step in providing information through ontology but how these classes are related to one another and what attributes does a class have gives the meaning to an ontology. W3C defines property in a much more clear way. The way it works is that each property has a domain that links it to a class and a range to specify what values that the property can take (17). From the naming conventions given in the guide, each property will start with a small case letter and each '-' is used as separator. There are two types of properties - object and data type.

#### Data Type Properties

Not each term necessarily has to be a class. Some information can be provided using data type properties that describe some attribute of a class. For each data type property a name and its type has to be set. The range for a data type properties is one of the XML defined data types, such as string, int, date, boolean and so on.

AI system is assigned a unique ID in the database and hence this property is named as has-id having 'int' as data type since it contains only integers. 'Ai\_system' also has a property called has-name that denotes the name of the AI and the type is string since it contains only literals. Similarly has-incident will be used to provide a summary of the unethical event linked to the concerned AI. In addition, a reference to the AIAAIC repository which can be used to review the incident in detail is saved using has-link property with string as type. Lastly, 'Organisation' class has a property called has-owner-name that represents the name of the owner. Two blocks given in fig.4.4 shows how each property is defined using Chowlk notation.

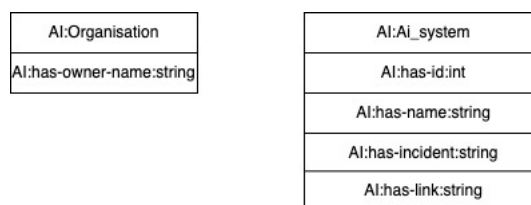


Figure 3.7: Data Properties in AIIA-Ontology defined using Chowlk

#### Object Properties

This kind of property is responsible for relating one class to another class. An object property has classes as range values. In AIIA-Ontology, the 'Ai\_system' class is related to the 'Organisation' class using has-owner object property. Therefore 'Ai\_system' is the domain and 'Organisation' is the range of has-owner property. Similarly other class relations are shown in the Figure 3.8 which also explicit the way Chowlk allows defining object properties i,e using the arrows. Other object properties include has-impact which

connects 'Ai\_system' and 'Impact\_zones' class. 'Risks' is the range of has-risks property with 'Ai\_system' being the domain. Lastly 'Values' and 'Principles' are related to the 'Risks' via has-violated-values and has-violated-principles object properties respectively.

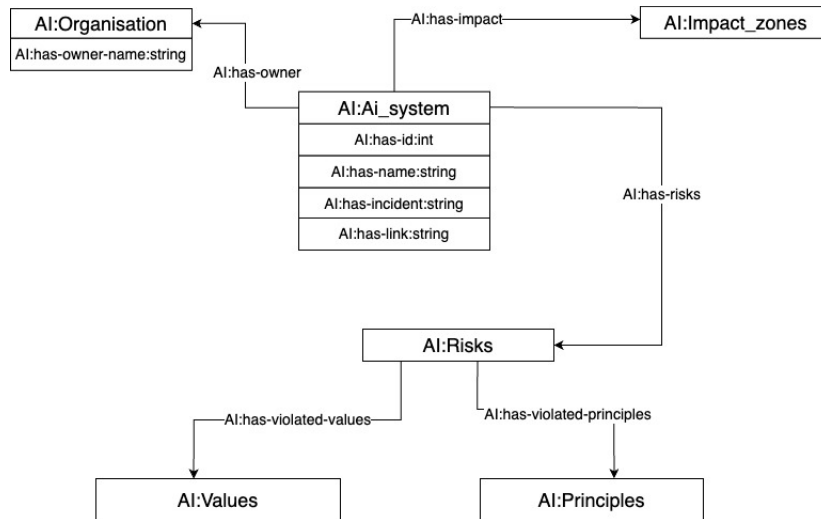


Figure 3.8: Object Properties of AIIA-Ontology defined using Chowlk

### 3.3.6 Restrictions

The purpose of impact assessment is to let the user know which values or principles have been violated by the concerned AI. In order to do so, ontology should be able to recognise by itself which particular risk will impact certain values and principles. To equip ontology with this knowledge a very special and useful element called restrictions are used. W3C (17) explains restrictions as a 'special kind of class' which describes an unknown class and they are always applied on properties. There are two kinds of restrictions. The first one is about cardinality that set a limit to the range of values a property can have. Another one which is used here is value constraints restriction. It allows setting some specific classes as range of a property so that no other class satisfies this restriction. It is important to understand that this kind of range is different from the general object property ranges.

There are various kinds of value constraints however the one that is used here is allValuesFrom. It is used to strictly tell that the values of a property can only and only be from the specified class in the range. It will be more clear with an example shown in fig 3.9 where a allValuesFrom restrictions is defined using Chowlk.

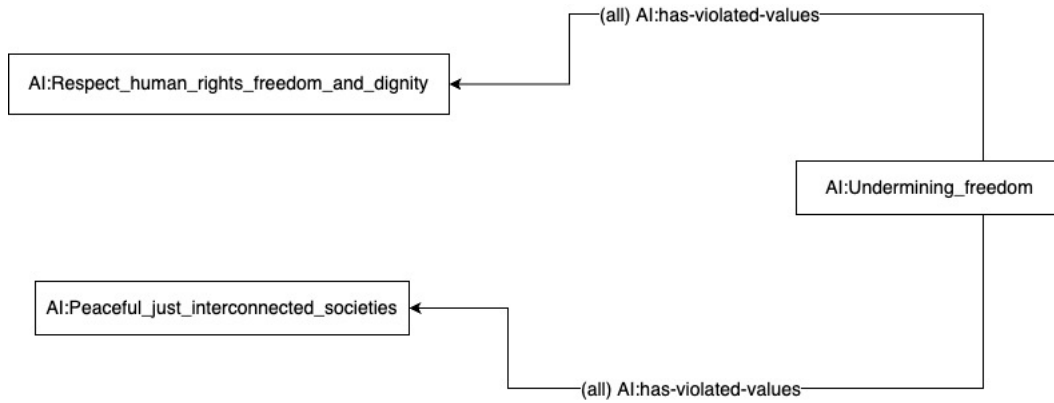


Figure 3.9: Defining restrictions using Chowlk

The significance of this restriction in fig.3.9 is that it tries to give information about the risk class called '*Undermining\_freedom*' linked to a property called *has-violated-values* that only accepts class '*Respect\_human\_rights\_freedom\_and\_dignity*' as well as '*Peaceful\_just\_interconnected\_societies*' as its range. This was just an illustration of one risk class, similarly all the risks are associated with one or more values and principles. Properties where the restriction is applied are *has-violated-values* and *has-violated-principles*. Instead of showing each one of them separately, table 3.3 is given below summarises each risk and its violations of specific values and principles.

Table 3.3: Risks and their associated values and principles

Subclass of 'Risks'	Values and principles getting violated
Harming_humans	Respect_human_rights_freedom_and_dignity
Undermining_freedom	Respect_human_rights_freedom_and_dignity
	Peaceful_just_and_interconnected_societies
Harming_environment_and_ecosystem	Environment_and_ecosystem_flourishing
Privacy_data_concerns	Right_to_privacy_and_data_protection
Benefits_not_shared	Fairness_and_non_discrimination
Raising_hatred_and_threats	Peaceful_just_and_interconnected_societies
Goals_and_objectives	Proportionality_and_do_no_harm
Lacking_human_intervention	Proportionality_and_do_no_harm
	Human_oversight_and_determination
Safety_security_issue	Safety_and_security
Lacks_transparency	Transparency_and_explainability



### 3.3.7 Individuals or Instances

Each subclass of '*Risks*' covers a wide range of issues under its umbrella. An Individual is a member or an instance of a particular class. Instances are the concepts at the highest granularity level (12). This key element is used to fill every subclass with the most specific risks. It will help in better refinement of AIIA-Ontology and get a deeper coverage of all the concerns identified by UNESCO (see table 3.1 and table 3.3). Fig 3.10 displays every subclass together with individual risks. Notice that an instance called 'segregating\_objectifying' is declared in two subclasses namely '*Harming\_humans*' and '*Raising\_hatred\_and\_threats*'. Ontologies allows defining instances belonging to more than one class.

## 3.4 Ontology Encoding

With the help of Chowlk all the concepts, properties and instances have been added in the ontology. However it is in the form of a diagram and in order to let machines understand, it needs to be formally coded. The diagrams.net has the option to download the diagram in XML which in turn gets converted into OWL using Chowlk's converter (29).

Now that the key elements of the ontology are present, the OWL file can be opened in any ontology editor like Protege to add other details like comments in order to add annotations that provide some descriptions for each terms. The first draft of AIIA-Ontology is ready at this point.

## 3.5 Querying Ontology

Ontologies can be queried to retrieve the information it has. SPARQL is a query language based on RDF subject, predicate object triples. W3C has released a documentation for SPARQL which helps in learning how some information is retrieved from the ontology knowledge base (30). Just like SQL, SPARQL also has a select, insert and delete queries.

```
SELECT ?s ?p ?o
WHERE { ?s ?o ?p
}
```

where ?s ?o ?p are variables. Although this paper does not aim at providing SPARQL tutorial, but it is important to familiarise with those queries that helps in getting information relevant for impact assessment. Last but not the least all the queries shown in the next sub sections are executed in GraphDB, which is an online platform that supports loading ontology data and running SPARQL queries (31).

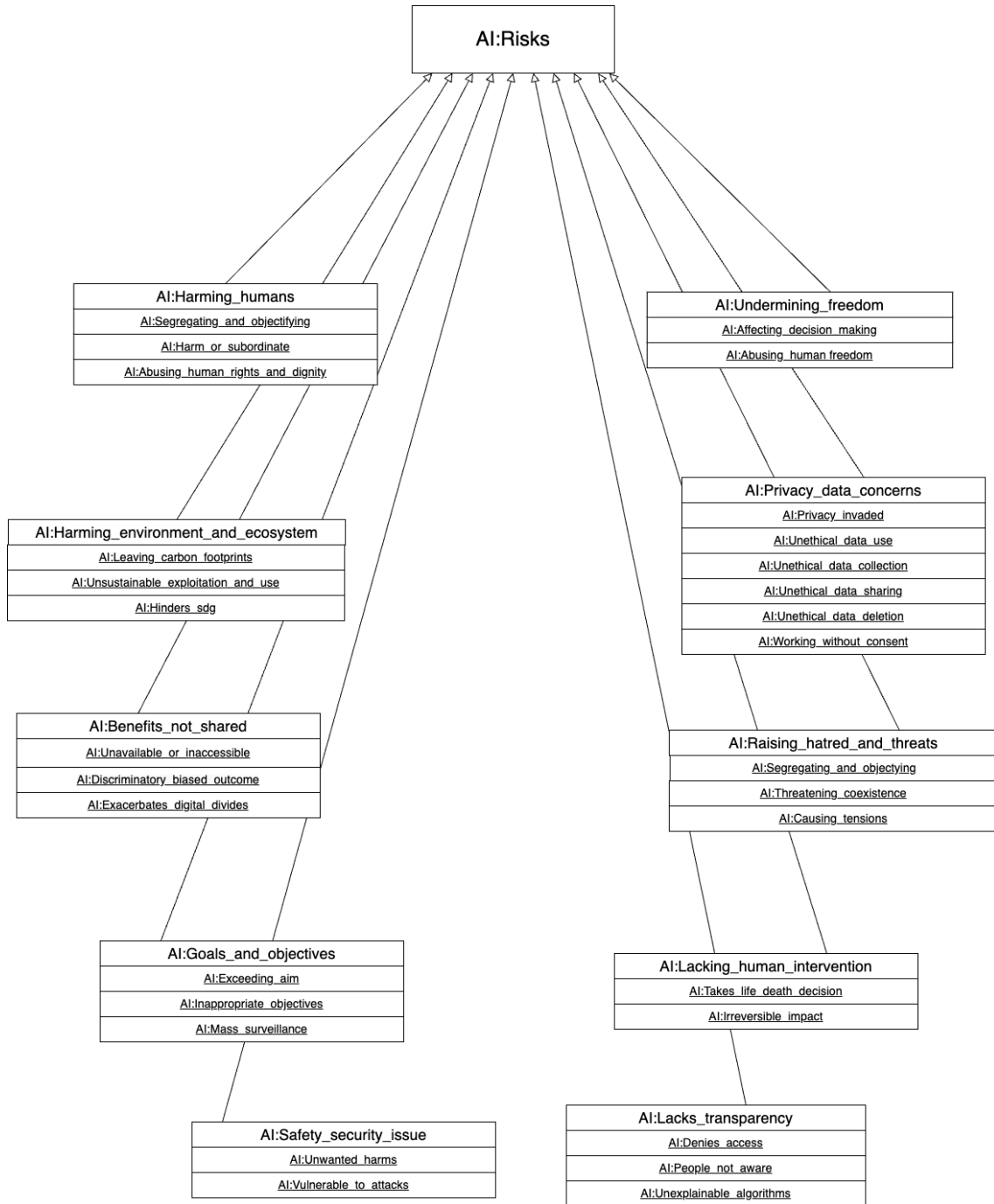


Figure 3.10: Instances of each subclass of 'Risks' defined in Chowlk

### 3.5.1 Query 1 : Finding all the subclasses of 'Risks'

This query helps in extracting all the risks that are going to be shown on the web application for selection. To retrieve the sub classes of 'Risks', `rdfs:subClassOf` axiom is used in the below format as this is how RDF language identifies subclasses.

```
SELECT ?risks
WHERE {
?risks http://www.w3.org/2000/01/rdf-schema#:subClassOf
<http://AIIA-Ontology/Risks>
}
```

The solution to the query is as shown in the fig 3.14. The results can be verified using the fig 3.6 where all the subclasses were displayed.

	risks	↕
1	<a href="#">AI:Benefits_not_shared</a>	
2	<a href="#">AI:Goals_and_objectives</a>	
3	<a href="#">AI:Harming_environment_and_ecosystem</a>	
4	<a href="#">AI:Harming_humans</a>	
5	<a href="#">AI:Lacking_human_intervention</a>	
6	<a href="#">AI:Lacks_transparency</a>	
7	<a href="#">AI:Privacy_data_concerns</a>	
8	<a href="#">AI:Raising_hatred_and_threats</a>	
9	<a href="#">AI:Safety_security_issue</a>	
10	<a href="#">AI:Undermining_freedom</a>	

Figure 3.11: Results of Query 1

### 3.5.2 Query 2 : Finding all the subclasses of 'Impact\_zones'

This query is similar to query 1 with only a slight change. The URI is changed to `<http://AIIA-Ontology/Impact_zones>`.

```
SELECT ?impacts
WHERE {
?impacts <http://www.w3.org/2000/01/rdf-schema#subClassOf>
<http://AIIA-Ontology/Impact_zones> }
```

The results can be verified by seeing the figure 3.3

	impacts
1	<a href="#">AI:Culture</a>
2	<a href="#">AI:Education</a>
3	<a href="#">AI:Media</a>
4	<a href="#">AI:Communities_and_groups</a>
5	<a href="#">AI:Countries</a>
6	<a href="#">AI:Environment</a>
7	<a href="#">AI:Employment_or_labour</a>
8	<a href="#">AI:Gender_equality</a>
9	<a href="#">AI:Health_care</a>
10	<a href="#">AI:Humans_privacy_safety</a>
11	<a href="#">AI:Humans_rights_freedom</a>
12	<a href="#">AI:Marginalised_or_vulnerable_people</a>

Figure 3.12: Results of Query 2

### 3.5.3 Query 3 : Finding all the instances of a given ‘Risks’

It will be used in the web form to let one select the risks that best describe the underlying concern. Instances are identified using `rdf:type` in RDF language. In addition to this the following query puts PREFIX syntax into use. It helps to avoid writing full URI each time.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX AI: <http://AIIA-Ontology/>
SELECT ?instances
WHERE ?instances rdf:type AI:Lacks_transparency
```

To check the results refer to fig 3.10

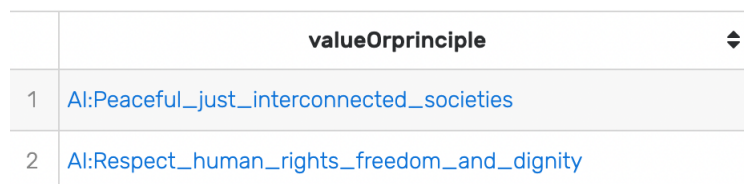
	instances
1	<a href="#">AI:Denies_access</a>
2	<a href="#">AI:People_not_aware</a>
3	<a href="#">AI:Unexplainable_algorithms</a>

Figure 3.13: Results of Query 3

### 3.5.4 Query 4 : Finding all the values or principles violated by a given instance of 'Risk'

This query is quite important as it leverages the idea of using ontology for impact assessment. It is also a tricky one. First the node of the concerned risk has to be found in the ontology, then using the owl:allValuesFrom restriction those values or principles are returned that are in the range of that node. To illustrate an example the following query is executed that seeks the violated values from the ontology provided the risk is Abusing\_Human\_Freedom.

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX AI: <http://AIIA-Ontology/>
SELECT ?valueOrprinciple
WHERE {
?riskNode owl:allValuesFrom ?valueOrprinciple.
AI:Abusing_human_freedom rdf:type ?riskNode.
}
```



	valueOrprinciple	
1	<a href="#">AI:Peaceful_just_interconnected_societies</a>	
2	<a href="#">AI:Respect_human_rights_freedom_and_dignity</a>	

Figure 3.14: Results of Query 4

### 3.5.5 Query 4 : Inserting data

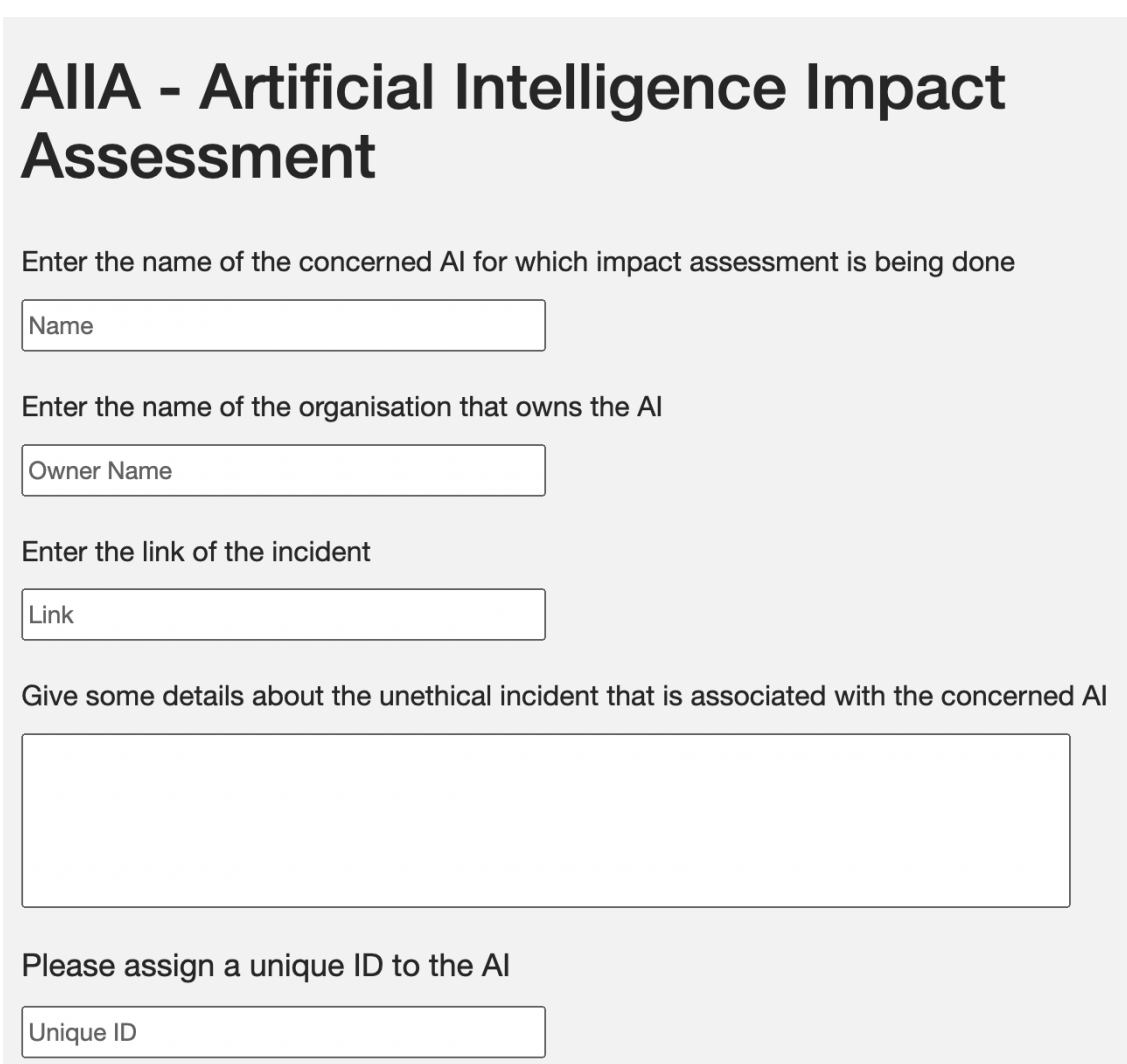
The database of an ontology can be extended by giving some values to each property that best describes the AI system. An example of insert query is shown below.

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX AI: <http://AIIA-Ontology/>
INSERT DATA {
AI:5 a AI:AISystem;
AI:has-name "Stanford Brainwash Dataset";
AI:has-id 50;
```

```
AI:has-incident "Stanford University data sharing";
AI:has-owner "Stanford University";
AI:has-impact AI:Culture;
AI:has-impact AI:Humans_privacy_safety.
AI:5 a AI:Risks;
AI:has-risk AI:Unethical_data_collection;
AI:has-risk AI:Unethical_data_sharing;
AI:has-violated-principles AI:Fairness_and_non_discrimination.
.}
```

## 3.6 Web form for External Impact Assessment

In section 3.3, the first draft of AIIA-Ontology was developed. It holds all the fundamental concepts extracted from UNESCO's recommendations. In section 3.5 it was shown how SPARQL queries are run on the ontology to retrieve some information. This section will see how real world AI controversies taken from AIAAIC repository are annotated using the vocabulary of AIIA-Ontology. Having an interface that allows choosing from a list of options will be convenient for impact assessment. Therefore to fulfil this last requirement a web form was designed using HTML and CSS to style the appearance. A python library called Flask<sup>1</sup> was used to set up a server on local machine. Fig 3.15 shows how the web form looks.



**AIIA - Artificial Intelligence Impact Assessment**

Enter the name of the concerned AI for which impact assessment is being done

Enter the name of the organisation that owns the AI

Enter the link of the incident

Give some details about the unethical incident that is associated with the concerned AI

Please assign a unique ID to the AI

Figure 3.15: Web form

As part of the external impact assessment, the web form requires one to enter details of the AI system such as name, owner, description of an incident, link or URL to cite the

---

<sup>1</sup>Python Flask

source if the incident (which is AIAAIC repository here). Then it asks user to choose impacts from the list drawn straight from the AIIA-Ontology using query 2 (see fig 3.16). Similarly it asks for the kind of risks that the AI has. All the options of the risks are also drawn directly from the AIIA-Ontology by running query 1 and 3. Flask handles the data exchange with HTML form. Python offers a library called SPARQLWrapper<sup>2</sup> that executes queries remotely and fetches results. GraphDB provides an endpoint for loading and storing ontology database. How it works is, SPARQLWrapper class sends a particular query to GraphDB server which contains ontology related data. GraphDB executes the query and sends back the results to SPARQLWrapper class which converts results in a python supporting format. Once the the web form is filled out completely, flask retrieves the data. Based on the selected risks, AIIA-Ontology sends back the associated values or principles that are violated. The working is better explained by taking an AI case scenario as an example and performing impact assessment on it.

Please select from the following options on which the AI has an impact

- Culture
- Education
- Media
- Communities and groups
- Countries
- Environment
- Employment or labour
- Gender equality
- Health care
- Humans privacy safety
- Humans rights freedom
- Marginalised or vulnerable people

Figure 3.16: List of Impact zones

### 3.6.1 Impact assessment on an incident involving AI

Zhihu<sup>3</sup> an online platform came into scrutiny when someone who got fired reported that the manager already knew about his plans of leaving the company. The resignation predictors are owned by Sangford Technologies which makes predictions by tracking the employees browser history and conversations with other employees.

---

<sup>2</sup>SPARQLWrapper Class

<sup>3</sup>Zhihu controversy



This incident was then described using the vocabulary of AIIA-Ontology by putting the following information in the web form.

**AI name** - Zhihu job resignation predictions

**Organisation who owns it** - Sangford Technologies

**AI incident description** - Predicts behaviour of employee by tracking history and conversation with other employees.

**AI has an impact on**

1. Employment or labour
2. Humans privacy safety

**Risk associated to AI**

1. Inappropriate objectives
2. Mass surveillance
3. Privacy invaded
4. Unethical data collection

This data is sent into the ontology database using insert data query. Also the impact assessment ends when AIIA-Ontology returns those values or principles that has been violated based on the selected risks. Refer fig ?? to see the final results. In this case study there has been violations of two principles.

1. Proportionality and do no harm
2. Right to privacy and data protection

## Here are the violations based on UNESCO's values and principles

Values :

Principles :

Proportionality and do no harm

Right to privacy and data protection

*Based on the following information provided by you*

### **AI name**

Zhihu job resignation predictions

### **Link to AI incident**

<https://www.aiaaic.org/aiaaic-repository/ai-and-algorithmic-incidents-and-controversies/zhihu-job-resignation-predictions>

### **AI Unique ID**

22

### **AI owned by**

Sangfor Technologies

### **AI incident**

Predicts behaviour of employee by tracking history and conversation with other employees.

### **AI has an impact on**

Employment or labour

Humans privacy safety

### **AI associated risks**

Inappropriate objectives

Mass surveillance

Privacy invaded

Unethical data collection

Figure 3.17: Results of external impact assessment

## 4 Evaluation

The entire external impact assessment is backed by AIIA-Ontology which extracts the concepts from the UNESCO recommendation document. If this ontology is implemented correctly, the aim of the paper can be fulfilled. Hence this chapter presents the evaluation of AIIA-Ontology using the following methods

- **Context Based Evaluation**

This will be used to verify that AIIA-Ontology has a good coverage of the UNESCO document with correct interpretation of the concepts.

- **Class Hierarchy Based Evaluation**

This is to validate the overall internal structure including class hierarchy, properties, etc.

- **Quality Based Evaluation**

This evaluation will check AIIA-Ontology for any common errors that yield in bad quality ontologies.

- **Application Based Evaluation**

AIIA-Ontology will be evaluated based on its performance when used for external impact assessment on a number of AI incidents taken from AIIAIC repository. Remember it was said that ontologies are finalised after going through multiple iterations. This evaluation step is repeated for three iterations to update and refine the ontology.

- **Database Evaluation**

This additional evaluation helps in verify the functioning of web application whether the database is maintained or not.

### 4.1 Context Based Evaluation

One of the methods stated in the paper by J. Raad (23) was comparing an ontology to some "gold standard" ontologies having similar motives. This will be the first step towards validating the overall coverage over UNESCO's recommendations. But as

mentioned earlier there are no other ontologies based on UNESCO. So alternatively an ongoing project called PATH-AI (32), led by researchers from University of Edinburgh and a Japan Science Institute, is examined to validate the idea behind AIIA-Ontology. These researchers are keen on exploring and extending the AI ethics discussions. One of the aims of PATH-AI project is to produce an ethical impact assessment (EIA). As of now the group has published a workbook<sup>1</sup> that gives an overview of UNESCO recommendations and how to operationalize them. They have recognised the work of UNESCO as the first ever approach taken from a global perspective. According to them, engaging all the stakeholders impacted by unethical AI helps in conducting EIA for it ensures that everyone's voice is heard while making policies and decisions. Although stakeholder engagement is not the scope of this project, but it is worth going through the steps as they tell how they utilised UNESCO's values and principles :

- **Step 1** : Describing the domain where the AI system will be used and the data on which it will be trained. This is to provide a broad description of the AI.
- **Step 2** : Identifying who is going to be affected. Apart from the environment and end users of AI, stakeholder groups can themselves be impacted. Some of the impacted fields are media, education, civil society, human rights institutions, communities or groups.
- **Step 3** : Identify which UNESCO values and principles are impacted. This step also includes reviewing the ethical concerns and risks or even benefits of AI systems.
- **Step 4** : Analysing the importance of each stakeholder which aims at identifying the most likely impacted stakeholder.

From the above steps, it is clear that the expert group is emphasising on identifying who may have been impacted (step 2), which of the UNESCO's standards have been impacted as well as identifying the risks associated with AI (step 3). This forms a good basis to state that AIIA-Ontology has been developed with a correct idea and interpretation as it identifies impact zones, followed by risks and the affected values and principles. Step 1 is although not exactly incorporated into the ontology however there is a class called Ai\_System that covers the high level information about the AI involved and the incident that it is held responsible for. Step 4 is not covered as it does not fit in with the scope of ontology. It rather supports the task of implementing stakeholder engagement which the research group is trying to achieve.

In addition to this, the PATH-AI project team also presents case study analysis to illustrate how a civil servant can do an assessment for an AI system dedicated to healthcare. The Fig4.1 is taken directly from the workbook which shows three questions that the

---

<sup>1</sup>The workbook pdf can be found in the supplementary material of this submission

civil servant should consider while assessing the AI.

REFLECTION QUESTIONS
1. What stakeholder groups, organisations and individuals do you think would be significantly impacted by this tool?
2. Could the introduction of this symptom-checking tool impact upon any of the UNESCO principles or values? If so, how might it help or hinder a health service which aims to uphold these principles and values?
3. What additional information might you need about the data collection, processing and sharing practices this company is using in order to assess its ethical impact?

Figure 4.1: AI assessment questions presented in Path-AI Workbook

It is interesting to see that the web form introduced as a part of external impact assessment in section 3.6 approaches exactly in the same manner. Similar to the first question in fig 4.1 the web form offers a list of options to identify the impacted zones by AI. Additionally the web form also lets one choose the risks that the AI owns. Finally based on the selection of risks a report is generated that tells which UNESCO standards are violated by the AI, thus satisfying the second question in fig 4.1. The only difference is that the case study in the workbook involves an AI that is in the development phase and the organisation is willing to fix the underlying issues. It is more of an internal impact assessment. This is the reason that third question in fig 4.1 involves learning the way AI system works. Whereas the purpose of AIIA-Ontology is to facilitate external impact assessment, meaning assessing an AI that is already in the market and has some unethical behaviour. Hence risks are identified and based on those certain violations are concluded.

From the above discussion, it can be concluded that AIIA-Ontology has successfully incorporated the desired concepts like impact areas, risks and UNESCO values and principles. The set of questions the web-form presents to conduct external impact assessment are also valid and relevant.

Moreover, AIIA-Ontology provides a mechanism to decide what kind of violations have occurred as risk is linked to certain values and principles. It therefore automates the generation of results and reduces human efforts. Due to this reason it can be said that AIIA-Ontology proves to be an effective model for external impact assessment, since it automatically gives users the list of violated values and principles provided the underlying risks.

## 4.2 Class Hierarchy Based Evaluation

Traditionally ontologies are manually coded using RDF/XML or OWL language, but to create AIIA-Ontology a tool called Chowlk was used which comes with predefined templates for expressing various elements of ontology (28). The ontology development process indeed got a lot easier since the tool allows one to define classes, properties, instances, etc exactly like any UML diagram. Gradually the ontology got populated with loads of classes and subclasses, the diagram became huge and complex. Although Chowlk converter (29) rendered the designed diagram into OWL language, it still needs a careful inspection to make sure every element of the ontology is present and well connected. This evaluation method will help in verifying the structure and syntax of ontology which was recommended by Gómez-Pérez (22)

The OWL file that was generated using Chowlk was loaded in Protege application in order to identify any element that might be missing or incorrectly positioned. Fig 4.2. displays how the window looks like when the AIIA-Ontology is opened in Protege. Every ontology contains a class called '*Thing*' which encloses the rest of the classes. Even an empty ontology will have a '*Thing*' class. Fig 4.2 also contains pictures that show '*Risks*', '*Impact\_zones*', '*Values*' and '*Principles*' encapsulating their respective sub-classes accurately.

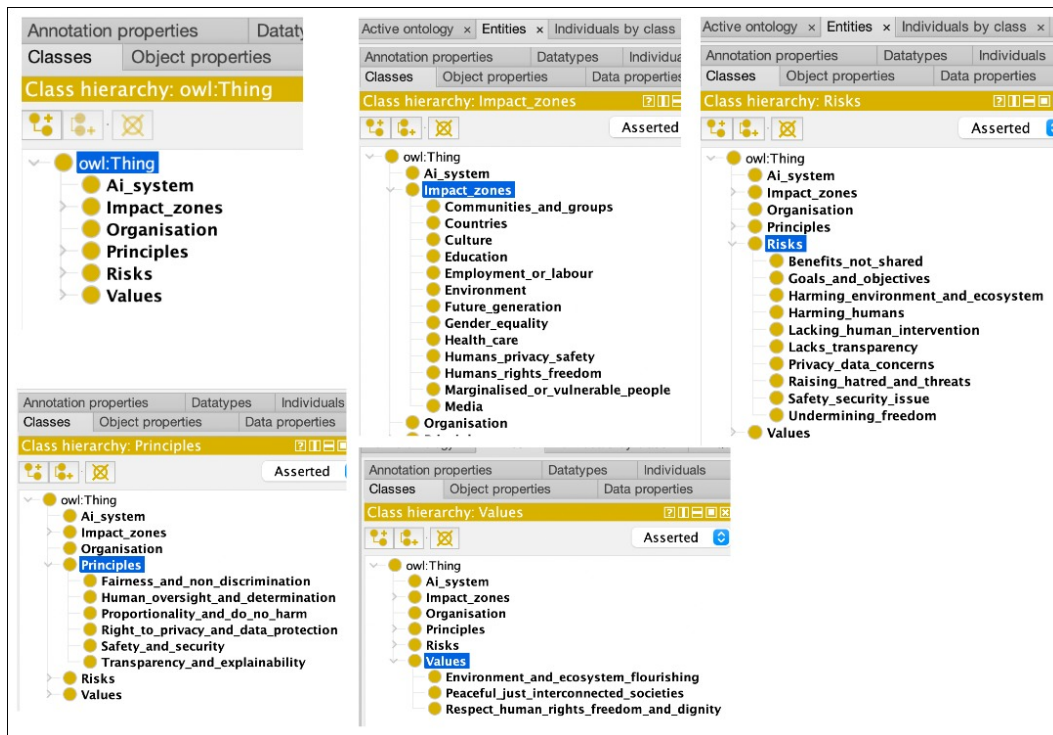


Figure 4.2: AIIA-Ontology classes in Protege application

To check whether all of the instances or individuals are present, Protege provides a tab

called Individuals, containing a list of all the instances belonging to the ontology (refer fig. 4.3). The picture on the right hand side shows how one can select a class in Protege to see its individual members along with the classes that it is a subclass of. Annotations such as label and comment are also visible for the selected class.

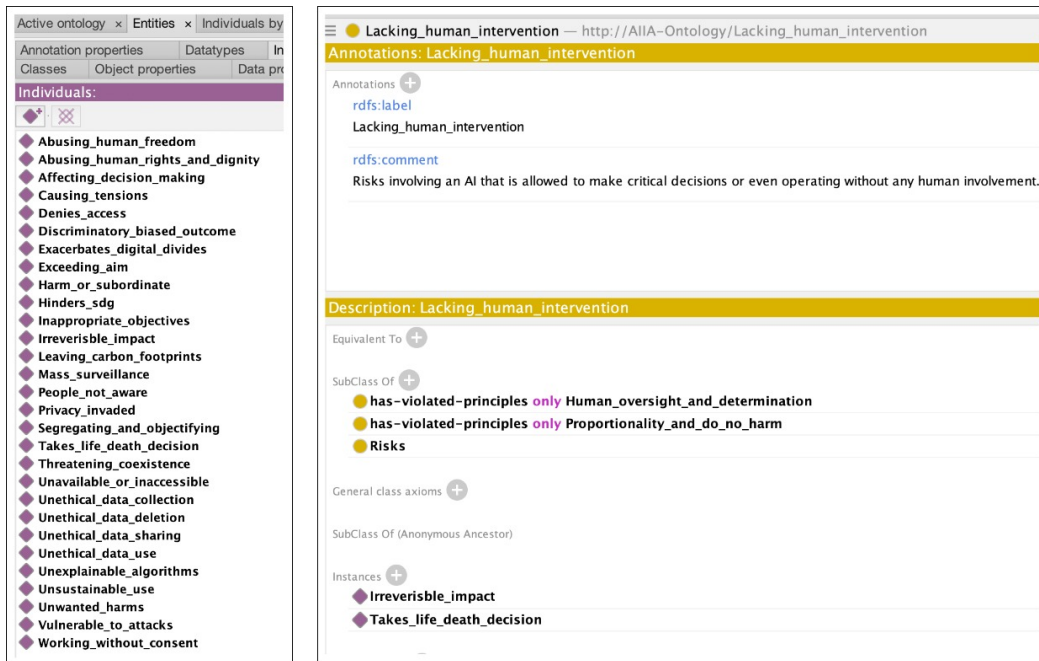


Figure 4.3: AIIA-Ontology instances in Protege

Looking at the fig. 4.4, it can be said that all the properties (both object and data) are correctly defined and assigned domain and range. Conclusively with this, class hierarchy and internal structure of AIIA-Ontology has been verified and it is confirmed that no element is missing or misplaced. It also confirms that Chowlk can exactly convert a diagram into an OWL language using correct notations (28) are obeyed.

### 4.3 Quality Based Evaluation

In the last section, AIIA-Ontology was evaluated for its overall structure. However, to claim reusability and reliability, having all the elements together is not enough. The developer needs to ensure that the ontology is meaningful and of great quality. A paper on common pitfalls in Ontology (24) lists twenty four most common practices that lead to poor quality. Each of these pitfalls is classified using one of the criteria which is affected - consistency, completeness and conciseness. OOPS <sup>2</sup> is an online platform that helps in detecting the presence of these common pitfalls. Fig 4.5 shows the results of OOPS evaluation over AIIA-Ontology. It also labels each pitfall depending on the level of importance it has in ontology modelling.

<sup>2</sup>OOPS validation

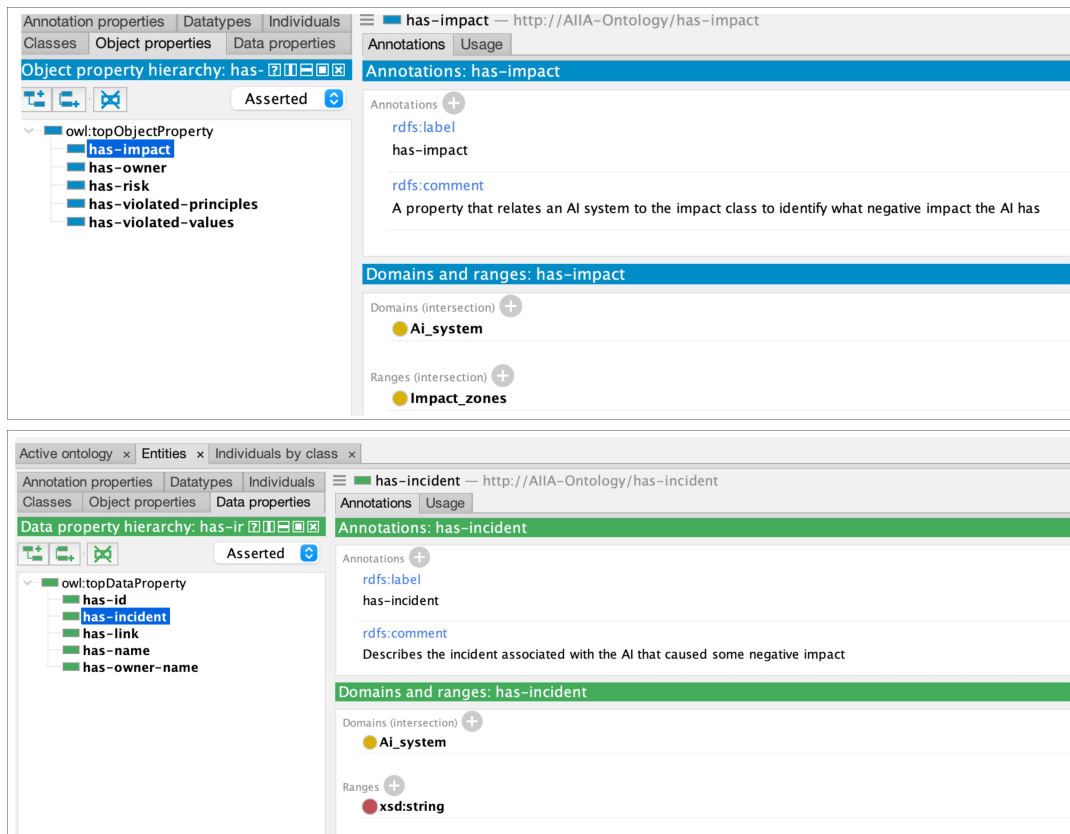


Figure 4.4: Object and data properties in Protege

Fig 4.5 tells there are a total of three pitfalls that have been detected, one of them labelled as important. However, there is no error that falls as critical since this category cannot be ignored. Even though these errors are not something that would disrupt the functioning of ontology, fixing them is considered a good practice to achieve high quality ontology. To make amendments in AIIA-Ontology, Protege is used from here.



It is obvious that not all the pitfalls are equally important; their impact in the

ontology will depend on multiple factors. For this reason, each pitfall has an importance level attached indicating how important it is. We have identified three levels:

**Critical**

It is crucial to correct the pitfall. Otherwise, it could affect the ontology consistency, reasoning, applicability, etc.

**Important**

Though not critical for ontology function, it is important to correct this type of pitfall.

**Minor**

It is not really a problem, but by correcting it we will make the ontology nicer.

## Evaluation results

P07. Merging different concepts in the same class	14 cases detected.	Minor
P10. Missing disjointness	ontology *	Important
P13. Inverse relationships not explicitly declared	5 cases detected.	Minor

Figure 4.5: OOPS validation results



### 1. Merging different concepts in the same class

Disjoint classes are the ones that do not share any instances (12). To address this issue each class is made disjoint with other classes. Protege has an option to specify disjoint classes. Fig 4.6 shows an example where 'Risks' class is declared disjoint to other classes.



Figure 4.6: Defining disjoint classes using Protege

### 2. Inverse relationships not explicitly declared

Apparently a good quality ontology expects the developer to create inverse relationships as well. Hence, for each object property an inverse property was defined that had the domain and range of the original one reversed. Table below shows the original properties and their corresponding inverse properties.

Object Property	Corresponding Inverse Property
has-impact	impacted-by
has-risks	associated-with
has-violated-values	values-violated-by
has-violated-principles	principles-violated-by
has-owner	owner-of

### 3. Merging different concepts in the same class

This error emerges because there are some elements that contain the words 'and' or 'or' (fig 4.7). This pitfall, although not really an issue, can be fixed by removing the words 'and' 'or' from all names.

#### Evaluation results

**P07. Merging different concepts in the same class** 14 cases detected. Minor

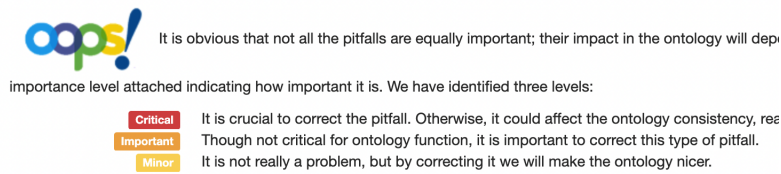
A class whose name refers to two or more different concepts is created.

This pitfall affects to the following ontology elements:

- "http://AIIA-Ontology/Respect\_human\_rights\_freedom\_and\_dignity"^^xsd:anyURI
- "http://AIIA-Ontology/Marginalised\_or\_vulnerable\_people"^^xsd:anyURI
- "http://AIIA-Ontology/Raising\_hatred\_and\_threats"^^xsd:anyURI
- "http://AIIA-Ontology/Safety\_and\_security"^^xsd:anyURI
- "http://AIIA-Ontology/Employment\_or\_labour"^^xsd:anyURI
- "http://AIIA-Ontology/Harming\_environment\_and\_ecosystem"^^xsd:anyURI
- "http://AIIA-Ontology/Goals\_and\_objectives"^^xsd:anyURI
- "http://AIIA-Ontology/Transparency\_and\_explainability"^^xsd:anyURI
- "http://AIIA-Ontology/Environment\_and\_ecosystem\_flourishing"^^xsd:anyURI
- "http://AIIA-Ontology/Right\_to\_privacy\_and\_data\_protection"^^xsd:anyURI
- "http://AIIA-Ontology/Communities\_and\_groups"^^xsd:anyURI
- "http://AIIA-Ontology/Proportionality\_and\_do\_no\_harm"^^xsd:anyURI
- "http://AIIA-Ontology/Fairness\_and\_non\_discrimination"^^xsd:anyURI
- "http://AIIA-Ontology/Human\_oversight\_and\_determination"^^xsd:anyURI

Figure 4.7: Merging different classes pitfall

After making the above changes, AIIA-Ontology was again tested by OOPS. This time no pitfalls were detected (see fig 4.8). Hence it can be stated that the ontology is of high quality as it passed the test for consistency, completeness and conciseness.



## Congratulations! OOPS did not find a single pitfall

Figure 4.8: No error detected

## 4.4 Application Based Evaluation

No evaluation of a model is complete without putting it to use in an application. The results of the application will help in drawing conclusions about the performance of the ontology in terms of the extent to which it fulfils a given task. The purpose of AIIA-Ontology is to conceptualise information in such a way that an external ethical impact assessment can be carried out on some AI incidents. Therefore, using the Web application developed in the section 3.6 real world case scenarios involving an AI system, provided in AIIAIC repository are annotated using AIIA-Ontology vocabulary.

As stated earlier, ontologies are finalised after a number of iterations. Keeping this in mind, the performance of AIIA-Ontology was evaluated over three iterations, where each iteration covers ten controversial AI from the AIIAIC database. It is important to understand that at this stage the ontology performance is measured from a developer point of view. Each complete assessment is given one of these labels

- **Satisfied**

This denotes that the developer is completely satisfied with the categories in the web form as well as the final results of the assessment which is the list of particular values and principles getting impacted.

- **Not Sure**

This code denotes that there was some confusion and the developer is not confident about both the selections and the final results.

- **Not Satisfied**

This denotes that the developer was not satisfied with the selections and did not get the expected results.

Although at this point the feedback is from the developer’s perspective and hence readers’ opinions about an AI incident can contradict. But once the web application is available to the public, better suggestions can be drawn. Next subsections show the results of the assessments in each iteration. Those AI incidents that are labelled as ‘Not Satisfied’ will only be taken into account for the improvement of the ontology. However it is essential to understand that ontologies are highly subjective. It means that the decision of making any updates entirely depends on the designer and also on the scope of the ontology.

#### 4.4.1 First Iteration

After conducting external impact assessment for 10 AI incidents, it was learnt that two of the incidents have been labelled as ‘Not Satisfied’ rest all are labelled ‘Satisfied’ (see fig. 4.9). Therefore two incidents will be considered for updating.

id	incident	values	principles	Label
1	Tesla crashing into private jet after being summoned by owner			Not Satisfied
2	Gives different scores to same person in different outfits		Fairness_and_non_discrimination	Satisfied
3	Too intrusive and maybe biased against darker skins		Fairness_non_discrimination	Satisfied
4	sentiment analysis on the basis of zoom meeting		Fairness_non_discrimination	Satisfied
5	Capturing videos without consent and sharing data		Right_to_privacy_data_protection	Satisfied
6	Manipulating its search algorithms in its own favour		Proportionality_do_no_harm	Satisfied
7	Video recorded without consent and data sharing		Transparency_explainability Right_to_privacy_data_protection	Satisfied
8	inaccurate sentencing that is unfair and biased racially	Respect_human_rights_freedom_dignity	Fairness_non_discrimination	Not Satisfied
9	Facial recognition and data processing without consent		Transparency_explainability Right_to_privacy_data_protection	Satisfied
10	collecting and processing data without consent		Transparency_explainability Right_to_privacy_data_protection	Satisfied

Figure 4.9: External Impact assessments results from iteration 1

- Tesla Crashing into private jet after being summoned by owner<sup>3</sup>**  
 In this scenario, AIIA-Ontology failed in finding out the impacted values and principles. This was due to the reason because the ontology does not talk about impacts on any property or object. Reflecting on the UNESCO recommendations, their only concerns are more centred around humans or on the environment. Property damage does not come under UNESCO domain of interest therefore no changes are made in the original ontology.
- Inaccurate sentencing that is unfair and biased racially<sup>4</sup>**  
 This incident is definitely unethical in the eyes of UNESCO. After studying the incident it was found that the AI was making decisions without human intervention. The ontology pointed out the values and principles being violated but it is

<sup>3</sup>Tesla car incident

<sup>4</sup>Inaccurate and biased sentencing

still labelled as 'Not satisfied' because according to UNESCO's 'Human oversight determination' principle must also be triggered. This failure happened because the '*Lacking\_human\_intervention*' class that is linked to this concerned principle in AIIA-Ontology does not have any instance that specifies this particular risk. Considering this, a new instance called '*No\_human\_intervention*' is defined.

#### 4.4.2 Second Iteration

The first iteration saw a change in the AIIA-Ontology. Fig 4.10 displays the results of next 10 AI incidents assessment. This time a total of four entries were labelled as 'Not Satisfied'. After studying each one of them, it was observed that all these AI lacked accuracy and so were not reliable. Reliability aspect can be considered but under which subclass of '*Risks*' it belongs to requires some brainstorming. Generally inaccurate AI causes some unwanted harm which in turn violates 'safety and security' principle. Therefore it seems logical to define a new instance called 'unreliability' inside the '*Safety\_security\_issues*' class.

id	incident	values	principles	Label
11	inaccurately generating customers background			Not Satisfied
12	Airbnb banning people without giving reasons and vague statements about user profiling		Transparency_explainability Proportionality_do_no_harm	Not Sure
13	intrusive and unethical data uses also the app is unreliable		Right_to_privacy_data_protection Proportionality_do_no_harm	Not Satisfied
14	spreading hate speech and disinformation	Peaceful_just_interconnected_societies		Satisfied
15	spreading misinformation and disinformation about Ukraine Russia war	Peaceful_just_interconnected_societies		Satisfied
16	Uber car crashing into woman	Respect_human_rights_freedom_dignity	Safety_security	Not Satisfied
17	using depfakes to spread false and mis or disinformation			Not Satisfied
18	Attacking people using drone	Respect_human_rights_freedom_dignity Peaceful_just_interconnected_societies	Proportionality_do_no_harm	Satisfied
19	hacking approx 3 million of ETH		Safety_security	Satisfied
20	Tinder charging more money from people of age above 30 or gay or lesbian	Peaceful_just_interconnected_societies Respect_human_rights_freedom_dignity	Fairness_non_discrimination	Satisfied

Figure 4.10: External Impact assessments results from iteration 2

In the fig there was only one entry labelled 'Not Sure'.

#### 4.4.3 Third Iteration

This time the impact assessment results were impressive as all the incidents except one was labelled as 'Satisfied' (see fig. 4.11). This tells that the recent changes in the ontology rendered accurate impact assessment results.

id	incident	values	principles	Label
21	Using robotic dog to test temperature of homeless people	Respect_human_rights_freedom_dignity		Satisfied
22	Predicts behaviour of employee by tracking history and conversation with other employees.		Proportionality_and_do_no_harm Right_to_privacy_and_data_protection	Satisfied
23	Creating pornography content without consent using deepfake technology		Proportionality_and_do_no_harm Right_to_privacy_and_data_protection	Satisfied
24	the results are not accurate		Safety_security	Satisfied
25	Uber failing in giving clarity of the new algorithm that seems to be cutting payments of drivers		Fairness_non_discrimination Transparency_explainability	Satisfied
26	A driverless car pulls away after getting stopped by the police for not turning light on		Safety_security	Satisfied
27	Use of facial recognition without legal permission		Right_to_privacy_data_protection Proportionality_do_no_harm	Satisfied
28	Houthi drones attacked in oil factory in Abu Dhabi	Respect_human_rights_freedom_dignity	Right_to_privacy_data_protection Proportionality_do_no_harm	Not sure
29	Using facial recognition to track foreign journalist, migrant women, student	Peaceful_just_interconnected_societies Respect_human_rights_freedom_dignity	Fairness_non_discrimination Right_to_privacy_data_protection Proportionality_do_no_harm	Satisfied
30	voice recognitions to predict customers habits		Right_to_privacy_data_protection Proportionality_do_no_harm	Satisfied

Figure 4.11: External Impact assessments results from iteration 3

#### 4.4.4 Database Evaluation

This is an additional evaluation which verifies that the data is being stored from the web application to GraphDB where the database is stored. The following query was executed on GraphDB which asks the server to return the data of an AI whose id is 12. Figure 4.12 confirms that the data is being stored properly as all the property values have been returned with correct information.

```

PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX AI: <http://AIIA-Ontology/>
SELECT distinct ?p ?o
WHERE{
AI:12 ?p ?o.
}

```

	p	o
1	<a href="#">rdf:type</a>	<a href="#">AI:Risks</a>
2	<a href="#">rdf:type</a>	<a href="#">AI:AI_system</a>
3	<a href="#">AI:has-id</a>	*12**xsd:integer
4	<a href="#">AI:has-impact</a>	<a href="#">AI:Media</a>
5	<a href="#">AI:has-impact</a>	<a href="#">AI:Employment_labour</a>
6	<a href="#">AI:has-incident</a>	"Airbnb banning people without giving reasons and vague statements about user profiling"
7	<a href="#">AI:has-link</a>	"https://www.aiaaic.org/aiaaic-repository/ai-and-algorithmic-incidents-and-controversies/airbnb-user-trustworthiness-scoring "
8	<a href="#">AI:has-name</a>	"Airbnb user trustworthiness scoring"
9	<a href="#">AI:has-owner-name</a>	"Airbnb"
10	<a href="#">AI:has-violated-principles</a>	<a href="#">AI:Proportionality_do_no_harm</a>
11	<a href="#">AI:has-violated-principles</a>	<a href="#">AI:Transparency_explainability</a>
12	<a href="#">AI:has-risks</a>	<a href="#">AI:Inappropriate_objectives</a>
13	<a href="#">AI:has-risks</a>	<a href="#">AI:People_not_aware</a>

Figure 4.12: Query Results in GraphDB

# 5 Conclusion

## 5.1 Final discussion

This chapter will reflect upon each research question and discuss to what extent have they been achieved. The first question revolves around understanding to what extent an ontology can be used to represent UNESCO's recommendations on the ethics of AI. Second question was about learning how effective this ontology can be as a model for external impact assessment. It all started with an overview of the proposed recommendations by UNESCO in section 3.2. Some important impact areas and underlying concerns of AI were identified specifically from an external assessment point of view. During the course of this research, basic knowledge about ontologies and their applications were gained. An ontology referred as AIIA-Ontology (Artificial Intelligence Impact Assessment) in this report was then created by following each step suggested by ontology experts. Section 3.3 saw a complete explanation of the entire development process of the first draft ontology. Then in section 3.5 a few queries were tested on the draft AIIA-Ontology which showed how some information is retrieved from the ontology that might be relevant for the impact assessment purpose. Lastly, in section 3.6 a web application was presented that offers end users who wish to perform impact assessment, an interactive interface in the form of web form. An example of assessment was illustrated too along with the final results.

In reference to the first question, AIIA-Ontology successfully captures all the values and principles outlined by UNESCO. It has information about risks and impacts that an AI can have. Moreover certain values and principles were linked with each subclass of risks to show their violations. It was very convenient to organise these concepts in the form of classes and subclasses. On the other hand properties helped in defining relationships among concepts. To further support, the first evaluation method which was context based can provide more evidence. A comparison of AIIA-Ontology was done with a workbook released by a group of experts who are also implementing an ethical impact assessment model based on UNESCO's AI ethics model. It was learnt that their interpretation exactly resonates with the approach taken in this research.

They even listed out a few questions that were directed towards the core concepts in AIIA-Ontology, for instance, impacted groups, the risks that an AI could have and then checking which values and principles it violates. Without doubt, the AIIA-Ontology succeeds enormously in extracting and representing just the right amount of information from UNESCO recommendations documents.

Now to address the second question of this research, multiple external impact assessments were conducted during the application based evaluation method in section 4.4. Overall the ontology managed to examine most of the AI incidents correctly. At the end of third iteration 9 out of 10 incidents were labelled as 'Satisfied'. Ontology development is an iterative task and with each cycle it just keeps getting better. This was evident as the results of third iteration were better than the first and second. One thing to note here is that the assessment was done by only developer's point of view.

What makes AIIA-Ontology even better as a model is the use of restrictions that allowed each risk class to be linked with one or more UNESCO standards. This alleviates human efforts from manually searching the document to decide which values or principles have been violated. As a result any person who is not familiar with UNESCO guidelines can perform external impact assessment to generate a report.

## 5.2 Future Work

Ontologies are developed in an iterative way and by each iteration it gets better. There is always a scope of improvising an ontology. The most important work is to apply AIIA-Ontology for more AI incidents and make relevant updates wherever possible. For example, there could be more risks related to artificial intelligence or impact zones that can be included. More people, especially expert groups working in different sectors, should be involved in performing impact assessment. By doing so, their feedback can help in resolving the incidents that were labelled as 'Not Sure'.

The ontology can also be extended further to model concepts like policies and actions suggested by UNESCO for mitigating each risk. So by this the ontology would not only highlight the values but suggest appropriate actions to regulate the problem. In addition to this sub classes in '*Impact\_zones*' class should be populated with individual elements like the impacted laws, stakeholders involved, which categories of people, etc. For example '*Humans\_rights\_freedom*' class can have instances describing the groups or bodies that work for human rights and freedom as well as the main laws that are impacted.

Another piece of work which is quite important in order to make ontology accessible and readable to all, is to translate every term into different languages. This would



help in communicating the UNESCO framework to the larger public. At last a formal documentation can be created that explains each and every terms in AIIA-Ontology. Although the terms are quite self explanatory but having a document will increase the clarity of concepts.

### 5.3 Final Reflection

In this research, I was introduced to the idea of AI ethics and how public organisations are dealing with it. UNESCO approach was by far the most generic and can be applied universally. In addition to this, I also explored the use of ontologies and learnt how to built one from the beginning.

During the course of this research I faced many challenges. The most challenging one was analysing the UNESCO's document from an external impact assessment perspective. It took several readings of the entire document. UNESCO has proposed some values that were a bit overlapping with one other . Therefore choosing the right got a little difficult. Identifying what risks each value and principle are addressing was a crucial and difficult task. But in the end I was happy with the risks I selected.

Ontology development using Chowlk got a lot more interesting as I was able to visualise it like a diagram simultaneously. Understanding the use of properties and restrictions indeed helped in relating concepts with each other. Evaluation of the ontology was also very critical for this report as the total assessment is based on it.

# Bibliography

- [1] European Commission : Artificial Intelligence for Europe URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&from=EN>
- [2] AIIAIC Repository : <https://www.aiaaic.org/aiaaic-repository/ai-and-algorithmic-incidents-and-controversies>
- [3] Gill, A.S., Germann, S. Conceptual and normative approaches to AI governance for a global digital ecosystem supportive of the UN Sustainable Development Goals (SDGs). *AI Ethics* 2, 293–301 (2022). URL: <https://doi.org/10.1007/s43681-021-00058-z>
- [4] Jobin, A., Ienca, M., Vayena, E. The global landscape of AI ethics guidelines. *Nat Mach Intell* 1, 389–399 (2019). URL: <https://doi.org/10.1038/s42256-019-0088-2>
- [5] Cathy Roche, Dave Lewis, P.J. Wall Artificial Intelligence Ethics: An Inclusive Global Discourse? arXiv:2108.09959
- [6] Preliminary study on the Ethics of Artificial Intelligence by UNESCO : <https://unesdoc.unesco.org/ark:/48223/pf0000367823>
- [7] UNESCO recommendations on the ethics of Artificial Intelligence : <https://unesdoc.unesco.org/ark:/48223/pf0000381137>
- [8] 2030 Agenda for Sustainable Development by UN : <https://sdgs.un.org/goals>
- [9] Gruber, Thomas R.. "A translation approach to portable ontology specifications." *Knowledge Acquisition* 5 (1993): 199-220.
- [10] Arbanas, K. Čubrilo, M. (2015). Ontology in Information Security. *Journal of Information and Organisational Sciences*, 39 (2), 107-136. Retrieved from <https://hrcak.srce.hr/149616>
- [11] Uschold, Michael Grüninger, Michael. (1996). *Ontologies: Principles, methods and applications*. The Knowledge Engineering Review. 11.

- [12] Noy, N. McGuinness, Deborah. (2001). *Ontology Development 101: A Guide to Creating Your First Ontology*. Knowledge Systems Laboratory. 32.
- [13] W3C ontologies in semantic web: <https://www.w3.org/standards/semanticweb/ontology.html>
- [14] Kamel, Magdi Lee, Ann Powers, Edward. (2007). *A Methodology for Developing Ontologies Using the Ontology Web Language (OWL)*. 261-268.
- [15] Semantic web document by W3C : <https://www.w3.org/standards/semanticweb/>
- [16] K. Palaniamma, S. Vijayalakshmi, "Ontology Based Meaningful Search Using Semantic Web and Natural Language Processing Techniques," *ICTACT Journal on Soft Computing*, vol. 4, no. .01, pp. 662-666, 2013
- [17] W3C document on OWL language : <https://www.w3.org/TR/owl-ref/>
- [18] Pascal Hitzler, Krzysztof Janowicz, and Tania Tudorache. 2020. *Ontology engineering: Current state, challenges, and future directions*. *Semant. web* 11, 1 (2020), 125–138. <https://doi.org/10.3233/SW-190382>
- [19] Bennett, Mike. "The financial industry business ontology: Best practice for big data." *Journal of Banking Regulation* 14 (2013): 255-268.
- [20] Robert Stevens, Alan Rector, Duncan Hull (2010) *What is an ontology?*. *Ontogenesis*. <http://ontogenesis.knowledgeblog.org/66>
- [21] Protege tool <https://protege.stanford.edu/products.php>
- [22] Gomez-Perez, Asuncion. (1995). *Some ideas and examples to evaluate ontologies*. 299-305. 10.1109/CAIA.1995.378808.
- [23] Raad, Joe Cruz, Christophe. (2015). *A Survey on Ontology Evaluation Methods*. 10.5220/0005591001790186.
- [24] Villalón, María Poveda, Mari Carmen Suárez-Figueroa and Asunción Gómez-Pérez. "A Double Classification of Common Pitfalls in Ontologies." (2010).
- [25] AIRO: an Ontology for expressing AI Risks : <https://delaramglp.github.io/AIRO/AIRODocumentation/>
- [26] Agrawal, Vivek. "Towards the Ontology of ISO/IEC 27005: 2011 Risk Management Standard." *HAISA* (2016).

- [27] Herzog, Almut Shahmehri, Nahid Duma, Claudiu. (2007). An Ontology of Information Security. IJISP. 1. 1-23. 10.4018/jisp.2007100101.
- [28] Chowlk notations : <https://chowlk.linkeddata.es/notation.html>
- [29] Chowlk converter : <https://chowlk.linkeddata.es/>
- [30] SPARQL w3c URL <https://www.w3.org/TR/sparql11-query/>
- [31] GraphDB online platform : <https://graphdb.ontotext.com/documentation/10.0/about-graphdb.html>
- [32] PATH-AI Project : <https://path-ai.org/>